Desarrollo de Sistemas de Diálogo Oral Adaptativos y Portables:

Reconocimiento de Emociones, Adaptación al Idioma y Evaluación de Campo

Zoraida Callejas Carrión

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Granada

ugr Universidad de Granada

Desarrollo de Sistemas de Diálogo Oral Adaptativos y Portables: Reconocimiento de Emociones, Adaptación al Idioma y Evaluación de Campo

Zoraida Callejas Carrión

Memoria para optar al grado de Doctora en Informática
con mención de Doctorado Europeo

Dirigida por:
Dr. D. Ramón López-Cózar Delgado

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Granada

Granada, abril 2008

Editor: Editorial de la Universidad de Granada
Autor: Zoraida Callejas Carrión
D.L.:Gr. 1263-2008
ISBN: 9788469143667

Agradecimientos

Hace unos años entré en contacto con los sistemas de diálogo, con la suerte de hacerlo de manos de alguien que disfruta trabajando y que disfruta además trabajando con ellos. Desde entonces, en su entusiasmo y su laboriosidad he tenido un referente magnífico para realizar mi trabajo con ilusión. Si la función de un director de tesis es guiar y aconsejar, no puede haber mejor forma de hacerlo que animando a curiosear y desarrollar las ideas propias con libertad. Gracias, Ramón.

Y para trabajar con libertad es necesario salir de nuestro laboratorio, despegar los ojos de nuestra pantalla, conocer qué hacen los demás y de esa forma enriquecer nuestras propias aportaciones. Pasteur decía que aunque la ciencia no tenga nacionalidad, los científicos sí la tenemos. Quisiera dar las gracias a Jan Nouza y todos los miembros del Laboratoř počítačového zpracování řeči, por demostrarme que esto no es un impedimento y que para formar un equipo lo importante es la voluntad de trabajar para llegar a objetivos comunes. Gracias por haberme enseñado tanto, en lo profesional y lo personal, invitándome y acogiéndome en esa primavera que no olvidaré.

Del mismo modo, muchísimas gracias a Michael McTear. Recuerdo cuando al comenzar el doctorado comentamos mis primeras investigaciones. Me embarqué en la empresa de realizar esta tesis doctoral, en gran parte, debido a sus ánimos y su interés. Gracias por ofrecerme siempre valiosas opiniones y consejos que me han hecho mejorar, aprender y avanzar por buen camino.

No es menor la suerte que he tenido al compartir el día a día con personas estupendas. Me gustaría dar gracias a todos los compañeros del departamento que me han animado, apoyado en mi trabajo, felicitado por mis logros, y que en definitiva han hecho más alegre el trayecto recorrido. Especialmente, a los amigos de los laboratorios 1 y 2, incluidos los que ya no están y los que cruzaron “el charco” temporalmente, con los que he vivido los mejores momentos y bebido muchos “cafés”.

Hay otros cafés que, aunque a veces virtuales, me han rescatado del naufragio. Los de mis tablas de salvación, los de esas amigas que a pesar de la distancia siempre me hacen llegar su afecto. Gracias Bea, MªPaz y Vane, sin vosotras, que me comprendéis como nadie, nada habría sido lo mismo. También en la distancia un agradecimiento a Hynek, mi guía del país de Hrabal, que con su buen humor hace pequeños los escollos más grandes: no olvidaré tus bramborák salvadoras.

Vaya el mayor agradecimiento a mi única certeza ante la incertidumbre y seguridad en los momentos de indecisión: mi familia. A mis padres y a Clara por ayudarme y creer en mí por encima de todo y en cualquier momento. Gracias por aguantar estoicamente los efectos secundarios de la tesis, los días de "deadline", las jornadas agotadoras... Si saltara al ruedo, iría por ustedes. Gracias también a Carmen por endulzarme desde niña las largas tardes de inglés.

Finalmente, David, gracias por tu paciencia infinita y tu cariño, por recordarme que vale la pena y darme alas para seguir adelante. Gracias por allanarme el camino y darme la mano en su transcurso. Aunque te nombre el último, *in my life* ...

Resumen

La presente tesis doctoral describe el trabajo realizado en tres de las líneas más exigentes y prometedoras del área de los sistemas de diálogo oral: el reconocimiento de emociones no actuadas, la adaptación de reconocedores del habla entre idiomas y la evaluación de campo (no restringida a laboratorio) de sistemas, empleando criterios tanto "objetivos" como "subjetivos". La investigación descrita en la tesis constituye una aportación novedosa a lo que los expertos han definido como los mayores retos de investigación del área: la adaptatibilidad y portabilidad de los sistemas de diálogo oral.

En primer lugar, en cuanto al reconocimiento de emociones, se presenta un estudio detallado acerca de cómo calcular e interpretar coeficientes de fiabilidad para la anotación de corpus con emociones reales. Se propone una nueva aproximación eficiente que mejora considerablemente el reconocimiento de las emociones, tanto por parte del sistema como respecto a los niveles de acuerdo entre los anotadores humanos, mediante el uso de diferentes fuentes de información contextuales. Por una parte, el proceso de anotación se mejora consiguiendo valores de acuerdo entre anotadores cercanos al máximo alcanzable incluso tratándose de anotadores no expertos. Por otra, se propone un algoritmo para el reconocimiento de emociones que extrae automáticamente la información contextual en tiempo de ejecución, obteniéndose resultados que suponen una mejora del 40 % en comparación con el estado del arte.

En segundo lugar, la adaptación de reconocedores del habla a diferentes lenguas se ha realizado durante una estancia de tres meses en la Technical University of Liberec (República Checa). Como resultado de esta investigación, se presenta una aproximación eficiente en tiempo y esfuerzo para adaptar un reconocedor del habla a otros idiomas. En concreto un reconocedor del habla checa a un idioma que es acústicamente muy similar (eslovaco) y a otro con un origen completamente diferente (español). La precisión obtenida en el reconocimiento es de aproximadamente el 70 % para el español y del 80 % para el eslovaco en tareas que precisan grandes vocabularios (alrededor de 150.000 palabras).

En tercer lugar, se han llevado a cabo diversos estudios estadísticos sobre la evaluación de campo de los sistemas de diálogo oral, proporcionando nuevas evidencias empíricas sobre las relaciones entre los diferentes criterios de evaluación. El estudio incluye tanto parámetros objetivos como subjetivos, prestando especial atención a la satisfacción del usuario y al éxito de la tarea, estudiando el impacto de diferentes aproximaciones para la gestión del diálogo y del nivel de experiencia del usuario empleando el sistema, así como el nivel de colaboración de los usuarios durante la interacción.

Todas las propuestas de la tesis se han evaluado con sistemas de diálogo reales. Para cumplir esta finalidad se desarrolló el sistema de diálogo oral UAH. El sistema se puso a disposición del público (via telefónica) en junio de 2005, habiéndose grabado la totalidad de diálogos con el sistema. Se han anotado las llamadas recibidas durante un año utilizando criterios de evaluación estándares (por ejemplo la tasa de errores por palabra). Este corpus se ha ampliado con la anotación de emociones por parte de nueve anotadores no expertos, que etiquetaron cada intervención de los usuarios como "neutro", "enfado", "duda" o "aburrimiento". Tanto los métodos de reconocimiento de emociones como los estudios de evaluación propuestos en la tesis se han evaluado en la práctica empleando el corpus UAH. Con respecto a la adaptación entre idiomas, su evaluación se llevó a cabo utilizando el sistema MyVoice, desarrollado por la Technical University of Liberec. La traducción de sus comandos para posibilitar la interacción en español es otra de las contribuciones de la tesis.

Los resultados empíricos obtenidos con los sistemas de diálogo se han verificado rigurosamente, midiéndose su significatividad mediante diferentes estudios estadísticos. Los resultados de la investigación descrita se han publicado en conferencias y revistas de prestigio, tanto nacionales como internaciones; habiéndose presentado además mediante diversas ponencias, pósters y demostraciones internacionales.

Índice general

Índice de tablas

Índice de figuras

Consultado el interés de la ciencia que tratamos de cultivar, es preciso comenzar por exponer las dificultades que tenemos que resolver desde el principio (...) pues es imposible desatar un nudo si no se sabe la manera de hacerlo.

Aristóteles, Metafísica (Libro III)

1 Introducción

Debido a un rápido incremento en potencialidad y reducción en su coste, los ordenadores se han convertido en una parte esencial de nuestro día a día. Vivimos rodeados de numerosos dispositivos electrónicos que nos dan información y funcionalidades; y estamos interesados cada vez más en acceder a los mismos en cualquier momento y lugar y en nuestro idioma. Por tanto, se necesitan nuevas interfaces que establezcan medios de comunicación naturales, eficientes e intuitivos entre los seres humanos y las máquinas. Los sistemas de diálogo oral se han convertido en una alternativa sólida para dotar a los ordenadores de capacidades inteligentes de comunicación, puesto que el habla es el método de comunicación más natural y flexible entre seres humanos.

1.1 Sistemas de diálogo oral

Un sistema de diálogo oral (SDO) es un software que acepta lenguaje natural como entrada y produce una salida en lenguaje natural, estableciendo una conversación con el usuario. Para gestionar la interacción con los usuarios de forma exitosa, los sistemas de diálogo oral suelen llevar a cabo cinco tareas principales: reconocimiento automático del habla (RAH), procesamiento del lenguaje natural (PLN), gestión del diálogo (GD), generación de lenguaje natural (GLN) y conversión texto-habla (CTH). Estas tareas se implementan por lo general en módulos distintos, la figura 1.1 muestra la arquitectura modular clásica para el desarrollo de un sistema de diálogo.

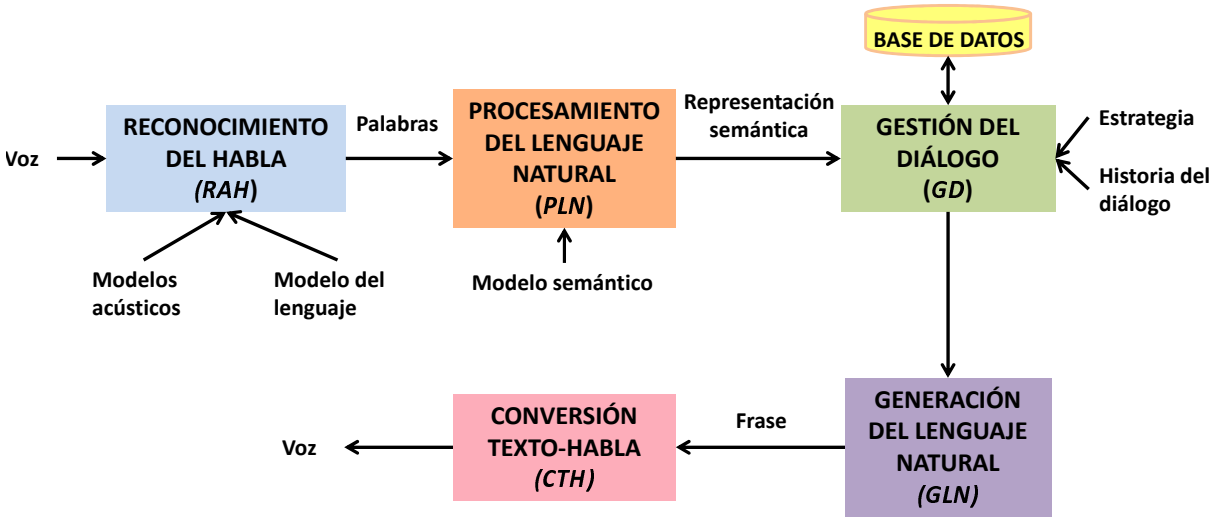


Figura 1.1. *Arquitectura modular de los sistemas de diálogo oral*

1.1.1. Reconocimiento del habla

El reconocimiento del habla es el proceso mediante el cual se obtiene el texto (frase pronunciada) correspondiente a un entrada oral pronunciada por un usuario. Se trata de una tarea muy compleja puesto que existe gran variabilidad en las características de la entrada, que cambian dependiendo de la lingüística de la frase, el locutor, el contexto de la interacción y el canal de transmisión. La variabilidad lingüística conlleva diferencias en los componentes fonéticos, sintácticos y semánticos que afectan a la señal de voz. La variabilidad interlocutor se refiere a la gran diferencia entre locutores según su forma de hablar, tono de voz, edad, sexo o nacionalidad. Incluso un mismo individuo no siempre pronuncia las mismas palabras de idéntica forma, puesto que las personas se ven afectadas por factores físicos y psicológicos que determinan la forma en que se comunican y que son altamente variables y por lo general impredecibles. Este fenómeno recibe el nombre de variabilidad intralocutor. Asimismo, las diferencias en los canales y/o dispositivos de comunicación también afectan a las señales de voz debido a los efectos derivados de la transmisión, como puede ser la reverberación. Finalmente, la variabilidad del entorno de interacción también es muy importante, ya que el reconocedor debe ser robusto frente a diferencias en el nivel de ruido ambiental.

Distintas tareas demandan diferentes complejidades del reconocedor de voz. En Cole et al. (1997) se enumeran ocho parámetros que permiten ajustar de forma óptima el reconocedor del habla: el modo de habla, el estilo de habla, la dependencia, el vocabulario, el modelo del lenguaje, la perplejidad, la relación señal-ruido y el dispositivo de entrada. En cuanto al modo de hablar, los reconocedores del habla pueden clasificarse en reconocedores de palabras aisladas o reconocedores de habla continua. Los primeros reconocen palabras separadas por pausas, mientras que los segundos son capaces de reconocer discurso natural en el que el locutor emplea su ritmo habitual. En cuanto al estilo de habla, ésta puede ser leída o espontánea. El habla espontánea es más natural, pero sin embargo posee ciertas peculiaridades como dudas o repeticiones que la hacen más difícil de reconocer. El reconocimiento del habla también puede ser dependiente o independiente del locutor. En el primer caso, los modelos acústicos se entrenan con la voz de un único locutor para el que se obtienen tasas de acierto óptimas; mientras que en el segundo caso, el reconocedor está preparado para reconocer a un amplio rango de locutores obteniendo tasas de acierto aceptables. Otro parámetro importante es el vocabulario aceptado, es decir, el número de palabras que el reconocedor del habla es capaz de reconocer. Por lo general, una aplicación se considera de gran vocabulario cuando supera las 5.000 palabras aceptadas (Jurafsky y Martin, 2000). Finalmente, para tratar el ruido y la variabilidad del canal, los reconocedores de voz emplean un modelo de canal ruidoso, es decir, abordan el problema del reconocimiento como si fuese necesario recuperar un mensaje que ha sido corrompido después de atravesar un canal con ruido. Para hacerlo, emplean modelos estocásticos que consideran los posibles mensajes originales y calculan cuál es más probable que sea el correcto.

1.1.2. Procesamiento del Lenguaje Natural

Una vez que el sistema de diálogo ha reconocido la frase pronunciada por el usuario, es necesario comprender su significado. El procesamiento del lenguaje natural permite obtener la semántica asociada a una cadena de texto. Generalmente realizar esta tarea conlleva el empleo de conocimiento morfológico, léxico, sintáctico, semántico, de discurso y pragmático. En la primera etapa, el conocimiento léxico y morfológico permite dividir las palabras en sus constituyentes, distinguiendo lexemas y morfemas: los lexemas

son la parte de las palabras que indica su semántica y los morfemas son los diferentes infijos y sufijos que permiten obtener las distintas clases de palabras. El análisis sintáctico obtiene la estructura jerárquica de las frases, sin embargo en el diálogo oral las frases frecuentemente se ven afectadas por las dificultades de lo que se denomina "disfluencia": pausas, repeticiones, estructuras sintácticas incompletas y correcciones (Gibbon et al., 2000). El análisis semántico extrae el significado de una estructura sintáctica compleja a partir del significado de sus constituyentes, mientras que en la etapa de procesamiento pragmático y de discurso, las frases se interpretan en el contexto del diálogo completo. La principal complejidad de esta última fase consiste en la resolución de anáforas y de ambigüedades derivadas de fenómenos como la ironía, el sarcasmo o el doble sentido.

Actualmente hay dos enfoques principales para abordar el problema del procesamiento del lenguaje natural: el enfoque basado en reglas y los modelos estadísticos aprendidos a partir de corpus de datos etiquetado. Los enfoques basados en reglas extraen información semántica basada en el análisis sintáctico-semántico de las frases, empleando gramáticas definidas para la tarea o detectando palabras clave con semántica asociada. En el caso de los métodos estadísticos, el proceso se basa en la definición de unidades de lenguaje con contenido semántico y el aprendizaje de modelos con muestras etiquetadas. Este tipo de análisis emplea un modelo probabilístico para identificar conceptos, marcadores y valores de los casos, con el fin de representar la relación entre los marcadores y sus valores, y decodificar semánticamente las frases de los usuarios (Minker, 1998).

1.1.3. Gestión del diálogo

No existe una definición globalmente aceptada acerca de las tareas que debe llevar a cabo un gestor del diálogo. Traum y Larsson (2003) sugieren que un gestor del diálogo debe cumplir cuatro funciones principales: i) actualizar el contexto del diálogo, ii) proveer de contexto en el que basar las interpretaciones, iii) coordinar al resto de módulos y iv) decidir que información dar al usuario y cuándo hacerlo. Por tanto, el gestor del diálogo tiene que tratar con distintas fuentes de información tales como los resultados del procesamiento del lenguaje natural, resultados de consultas a bases de datos, conocimiento acerca del dominio de aplicación, conocimiento acerca de los

usuarios y la historia previa del diálogo. Su complejidad depende de la tarea y la flexibilidad e iniciativa del diálogo. Bernsen et al. (1994) proporcionaron una taxonomía que muestra que para tareas pequeñas y sencillas un diálogo basado en palabras aisladas puede ser conveniente tanto con iniciativa dirigida por el usuario como dirigida por el sistema y que en este caso basta con una retroalimentación limitada. Sin embargo, para tareas grandes y bien estructuradas, se necesita que el diálogo sea dirigido por el sistema con una retroalimentación apropiada y realizando el seguimiento de la historia del diálogo empleando a su vez modelos de usuario sencillos. Para aplicaciones más grandes y mal estructuradas, son necesarios diálogos de iniciativa mixta con predicción automática, historia del diálogo y modelos de usuario avanzados.

La estrategia de gestión del diálogo más sencilla consiste en modelar el diálogo como una máquina de estados finitos en la que las transiciones entre las respuestas del sistema se determinan mediante las acciones de los usuarios. Dichas acciones son las respuestas de los usuarios al sistema, que se codifican en gramáticas de reconocimiento. Una extensión significativa son los enfoques basados en frames, desarrollados para solventar la falta de flexibilidad de las gramáticas de diálogo, constituyen el enfoque empleado por la mayoría de los sistemas comerciales actuales. Al contrario que la aproximación de estados finitos, los gestores de diálogo basados en frames no emplean un camino de diálogo predeterminado sino que usan una estructura compuesta de campos, uno por cada dato que el sistema deba obtener del usuario. De esta forma, el sistema interpreta el habla tanto para adquirir la suficiente información como para realizar una acción específica (hasta que rellena los campos). La ventaja de este enfoque es que permite capturar varios datos cada vez y que el usuario puede aportar la información en cualquier orden (se puede rellenar más de un campo en cada turno del diálogo y en cualquier orden).

Para dominios más complejos se emplea gestión del diálogo basada en planes, éstos se basan en la idea de que los seres humanos se comunican alcanzando objetivos y que durante la interacción pueden cambiar el estado mental de su interlocutor. Por tanto, los gestores de diálogo basados en planes modelan el diálogo como una cooperación entre el usuario y el sistema para alcanzar objetivos comunes. De esta forma, cada frase no se considera como una cadena de texto, sino como un acto del diálogo en el que el usuario comunica sus intenciones. Con cada intervención del usuario el sistema refina el modelo en el que intenta predecir sus intenciones y objetivos. Este

comportamiento se repite recursivamente para generar los turnos del sistema en respuesta a las frases del usuario hasta completar el objetivo de la tarea.

Este último enfoque está íntimamente relacionado con la teoría de diálogo denominada "del estado de información". El estado de información de un diálogo representa la información necesaria para distinguirlo unívocamente del resto de diálogos y que permite obtener un contexto sobre el que basar la siguiente intervención del sistema; por lo general esta información consiste en las intervenciones acumuladas del usuario y las acciones previas del diálogo. El estado de información a veces se ha denominado también almacén de conversación, contexto del discurso o estado mental. Según la teoría de estado de información, las principales tareas de un gestor de diálogo son actualizar el estado de información y seleccionar la siguiente acción del sistema basándose en las acciones del usuario observadas. El proyecto Trindi (TRINDIConsortium, 2001), propuso una arquitectura y desarrolló un toolkit para la construcción de gestores del diálogo basándose en este enfoque.

Por otra parte, cuando es necesario ejecutar y monitorizar operaciones en un dominio de aplicación que cambie dinámicamente, se puede emplear un enfoque basado en agentes. La aproximación basada en agentes para la gestión de diálogo posibilita combinar los beneficios de distintos modelos de gestión, como por ejemplo el de estados finitos y el de frames (Chu et al., 2005); así como de diferentes iniciativas de gestión de diálogo, como la dirigida por el sistema y la iniciativa mixta (Walker et al., 1998a), que pueden emplearse alternativamente de forma adaptativa. Además, permite combinar enfoques basados en reglas y de aprendizaje automático (Turunen et al., 2004).

Por último, la aplicación de enfoques de aprendizaje automático al diseño de estrategias de gestión del diálogo es un área en crecimiento. Los enfoques de aprendizaje automático para la gestión de diálogo se basan en aprender estrategias óptimas a partir de un corpus de diálogos usuario-máquina empleando métodos de "ensayo y error" automatizados, en lugar de basarse en principios de diseño empíricos (Griol, 2007). El modelo de procesos de decisión de Markov (MDP) sirve en la mayoría de estos enfoques como una representación formal del diálogo entre persona y máquina y proporciona una base para formular los problemas de aprendizaje de estrategias (Cuayáhuitl et al., 2006; Lemon et al., 2006; Williams y Young, 2007).

1.1.4. Generación del Lenguaje Natural

La generación del lenguaje natural es el proceso de obtención de textos en lenguaje natural a partir de una representación no lingüística. Se suele llevar a cabo en cinco pasos: organización del contenido, distribución del contenido en frases, lexicalización, generación de expresiones referenciales y realización lingüística. Es importante obtener mensajes legibles, optimizando el texto empleando expresiones referenciales y nexos y adaptando el vocabulario y la complejidad de las estructuras sintácticas a la destreza lingüística del usuario.

El enfoque más sencillo consiste en emplear mensajes de texto predefinidos (como por ejemplo mensajes de error o avisos). Aunque es intuitivo, este enfoque carece de flexibilidad. El siguiente nivel de sofisticación es la generación basada en plantillas, en las que una misma estructura de mensaje se utiliza incluyendo ligeras alteraciones. El enfoque de plantilla se emplea principalmente para la generación de frases en aplicaciones con texto de estructura muy regular como informes de negocios.

Los sistemas basados en frases emplean lo que puede considerarse como plantillas generalizadas, a nivel de oración (las frases se asemejan a reglas gramaticales), o a nivel del discurso (en este caso a menudo se denominan planes de texto). En estos sistemas se selecciona en primer lugar un patrón para emparejar el nivel superior de la entrada y seguidamente cada parte del patrón se amplía en uno más específico que se empareja con una determinada porción de la misma. El proceso en cascada se detiene cuando cada patrón ha sido substituido por una o más palabras.

Finalmente, los sistemas basados en características representan, en cierto modo, el nivel máximo de generalización y flexibilidad. En estos sistemas, cada alternativa mínima posible de expresión se representa por una sola característica; por ejemplo, si la frase es positiva o negativa, si es una pregunta o un imperativo o una declaración, o su tiempo verbal. Para ordenar las características es necesario emplear conocimiento lingüístico, alternativamente se puede generar el lenguaje natural basándose en corpus (Oh y Rudnicky, 2000), construyendo las elocuciones del sistema de forma estocástica.

1.1.5. Síntesis de texto a voz

Los sintetizadores de texto a voz transforman un texto en una señal acústica. Un sistema de síntesis oral se compone de dos partes: un "front-end" y un "back-end". El front-end realiza dos tareas fundamentales. En primer lugar, convierte el texto plano, (que contiene símbolos tales como números y abreviaturas) en sus palabras asociadas. Este proceso se denomina usualmente normalización, proceso previo, o tokenización del texto. En segundo lugar, asigna transcripciones fonéticas a cada palabra, y divide y marca el texto en unidades prosódicas, es decir frases, cláusulas, y oraciones. El proceso de asignar transcripciones fonéticas a las palabras se llama conversión texto-a-fonema o conversión de grafema-a-fonema. La salida del back-end es la representación simbólica constituida por las transcripciones fonéticas y la información prosódica.

El back-end (denominado a menudo sintetizador) convierte la representación lingüística simbólica en sonido. Por una parte, la síntesis del habla se puede basar en la producción de voz humana, este es el caso de la síntesis paramétrica (que simula los parámetros fisiológicos del tracto vocal) y de la síntesis basada en armónicos (que modela la vibración de las cuerdas vocales). En esta última técnica, parámetros tales como la frecuencia fundamental, la expresión, y los niveles de ruido se varían en el tiempo para crear la onda del discurso artificial. Otra aproximación basada en modelos fisiológicos es la síntesis articulatoria, que comprende técnicas de cálculo para modelar el tracto vocal humano y los procesos de articulación.

Por otra parte, la síntesis concatenativa emplea unidades pregrabadas de voz humana. Se basa en la concatenación de fragmentos de voz pregrabados. Generalmente, produce la voz sintetizada más natural; sin embargo, las diferencias entre variaciones naturales en el habla y la naturaleza de las técnicas automatizadas para segmentar las señales acústicas dan lugar a veces a interferencias audibles en la salida. La calidad de la voz sintetizada depende del tamaño de la unidad de síntesis empleada. Suelen utilizarse grandes bases de datos de voz pregrabada donde cada elocución registrada se divide en alguno de los siguientes segmentos: fonemas individuales, sílabas, morfemas, palabras, frases, y oraciones. Existe un compromiso entre la inteligibilidad y naturalidad de la voz generada y la automatización del procedimiento de síntesis. Por ejemplo, la síntesis basada en palabras completas es más inteligible que la basada en fonemas, pero para cada nueva palabra es necesario obtener

una nueva grabación mientras que los fonemas permiten construir cualquier nueva palabra. Por un lado, la síntesis específica del dominio concatena palabras y frases pregrabadas para crear elocuciones completas. Se utiliza cuando la variedad de textos que el sistema debe generar como salida se limita a un dominio específico, como información de horarios o informes meteorológicos. Por otro lado, la síntesis de difonemas utiliza una base de datos mínima del habla que contiene todos los difonemas (transiciones de sonido a sonido) que ocurren en un idioma. El número de difonemas depende de la fonotáctica de la lengua: por ejemplo, el español tiene cerca de 800 difonemas, y el alemán cerca de 2.500. En la síntesis de difonemas, solamente es necesario incluir un ejemplo de cada difonema en la base de datos del habla.

Finalmente, la síntesis basada en HMM es un método en el que se modela simultáneamente el espectro de frecuencias (tracto vocal), frecuencia fundamental (fuente vocal), y duración de la voz(prosodia) utilizando HMMs. La voz se genera a partir de los mismos HMMs basándose en el criterio de máxima similitud.

1.2 Aplicaciones de los SDO

La complejidad de la interacción entre el usuario y el sistema de diálogo puede variar y algunos de los componentes previamente descritos pueden no ser necesarios. Por ejemplo, para un menú simple, no es necesario un análisis semántico. Sin embargo, para el desarrollo de un compañero conversacional deben utilizarse todos los módulos para interpretar la entrada del usuario, tomar decisiones justificadas sobre qué respuesta debe generar el sistema, y finalmente adaptar la respuesta a las necesidades y expectativas de usuario.

Existe un gran número de tareas que pueden desempeñarse mediante el uso de sistemas de diálogo oral. La más extendida es la recuperación de información. Algunos ejemplos de aplicación son los siguientes:

- DARPA Communicator - Interfaces conversacionales inteligentes para la reserva de billetes de avión (DARPA, 1992, 1994).

- Voyager - Información turística del área de Boston (Glass et al., 1995).

- ARISE - Sistema de información ferroviaria en distintos idiomas europeos (den Os et al., 1999).
- AUGUST - Sistema de diálogo sueco que emplea un agente animado para dar información acerca de Estocolmo (J. Gustafson y Lundeberg, 1999).
- Adapt - Sistema de diálogo multimodal para buscar apartamentos en el mercado inmobiliario de Estocolmo (Gustafson et al., 2000).
- Jupiter - Predicción meteorológica por teléfono (Zue et al., 2000).
- Mercury - Sistema de reserva de vuelos e información meteorológica (Seneff y Polifroni, 2000).
- CTT-Bank - Sistema de banca operado oralmente por teléfono (Melin et al., 2001).
- SmartKom - Sistema de diálogo multimodal alemán con diversos dominios de aplicación, entre ellos la reserva de entradas de cine (Alexandersson y Becker, 2001).
- Let's Go - Sistema de diálogo oral para obtener información acerca de los autobuses de Pittsburgh, orientao principalmente a extranjeros y ancianos (Raux et al., 2005).
- Amities project - Interacción multilingüe con servicios e información bancaria (Hardy et al., 2006).
- DisCoH - Sistemas de diálogo oral de ayuda en congresos (Andeani et al., 2006).
- HIGGINS - Navegación y guía de ciudades para peatones (Skantze et al., 2006).
- TALK TownInfo - Sistema de diálogo multimodal para escenarios de información turística (Lemon et al., 2006).
- Conquest - Sistema de diálogo oral que aporta información de horarios en congresos (Bohus et al., 2007).

Los sistemas de diálogo se han empleado también para la educación y el entrenamiento, particularmente con el fin de mejorar habilidades fonéticas y lingüísticas de los usuarios:

- LARRI - Sistema de diálogo multimodal que asiste y guia a personal de aviones de combate durante tareas de mantenimiento (Bohus y Rudnicky, 2002).
- ITSPOKE - Sistemas de diálogo oral de tutorización que conversa con los estudiantes para darles información y corregir sus errores de aprendizaje (Litman y Silliman, 2004).
- Radiobot-CFF - Sistema de diálogo oral que participa en conversaciones que ayudan a entrenar soldados en los procedimientos de inicio de misiones de artillería (Roque et al., 2006b).
- VOCALIZA - Sistema de diálogo para terapias de voz en el idioma español que ayuda a los terapeutas a entrenar a usuarios con distintas patologías y carencias en su habilidad lingüística (Vaquero et al., 2006).
- LISTEN - Tutor de lectura que muestra historias en una pantalla y escucha a los niños mientras leen en voz alta (Mostow, 2008).

La interacción oral puede ser la única manera de acceder a la información en muchos casos, como cuando la pantalla disponible es demasiado pequeña para mostrar la información (p.ej. en dispositivos de bolsillo) o cuando los ojos del usuario están ocupados en otras tareas (p.ej. en la conducción):

- MUST - Servicios de información multimodal y multilingüe para terminales móviles de tamaño reducido (Boves y Os, 2002).
- VICO - Virtual Intelligence CO-driver. Permite la interacción natural entre las personas y dispositivos digitales y servicios en el automóvil (Mattasoni et al., 2002).
- Athosmail - Sistema multilingüe y adaptativo para la lectura de correos electrónicos utilizando teléfonos móviles (Jokinen et al., 2004).
- CHAT - Asistente conversacional en tareas dentro del automóvil como por ejemplo operar un reproductor de MP3 (Weng et al., 2006).

- DICO - Sistema de diálogo multimodal que permite al conductor de un vehículo controlar dispositivos y acceder a servicios de Internet mediante la voz (Villing y Larsson, 2006).

La interacción oral es también útil para el control de dispositivos y robots, especialmente en entornos inteligentes:

- WITAS - Interfaz para sostener conversaciones multimodales con el robot-helicóptero WITAS (Lemon et al., 2001).
- ODISEA - Odisea es un sistema de diálogo oral que permite la interacción entre usuarios y entornos inteligentes. Los componentes del diálogo se generan automáticamente y permiten la interacción oral entre el entorno y los usuarios (Montoro et al., 2004, 2006).
- SENECA - Interfaz oral para aplicaciones de entretenimiento, navegación y comunicación en entornos móviles (Minker et al., 2004a).
- Clarissa - Navegador manejable mediante la voz, permitiendo a los astronautas ser más eficientes con sus manos y ojos y prestar atención completa a la tarea mientras que navegan con el sistema utilizando comandos orales (Rayner et al., 2005).
- MIMUS - Sistema de diálogo multimodal para controlar un hogar inteligente (Pérez et al., 2006).
- STanford AI Robot (STAIR) - Asistente conversacional robótico para la oficina y el hogar (Krsmanovic et al., 2006).
- Cogniron - El robot compañero cognitivo (Menezes et al., 2007).

Finalmente, los agentes virtuales y compañeros constituyen la aplicación más exigente de los sistemas de diálogo:

- Collagen - Asistentes conversacionales y agentes colaborativos (Rich y Sidner, 1998).
- AVATALK - Diálogos naturales e interactivos con seres humanos virtuales (Hubal et al., 2000).

- COMIC - Diseño de cuartos de baño mediante la interacción con un avatar capaz de generar expresiones faciales (Catizone et al., 2003).

- NICE - Personajes históricos y literarios capaces de comunicarse de forma natural y divertida con niños y adolescentes (Corradini et al., 2004).

1.3 Evolución de los SDO

Los seres humanos han imaginado siempre la posibilidad de comunicarse con seres artificiales. Existen muchos ejemplos en el cine y literatura, algunos de los más antiguos se pueden encontrar en la mitología griega y romana en las que los héroes conversaban con estatuas de divinidades o guerreros de bronce. Los primeros intentos serios de construir sistemas parlantes se remontan a los siglos XVIII y XIX en los que se construyeron los primeros autómatas que imitaban la conducta humana. Inicialmente, se trataba de máquinas en las cuales los maestros relojeros aplicaron toda su habilidad para construir animales o pequeños muñecos que podían producir sonidos. En 1770, el barón Von Kempelen desarrolló el primer autómata que generaba palabras enteras y frases cortas, que posteriormente fue mejorado por Josef Faber para construir la máquina Euphonia en 1857. Euphonia imitaba el mecanismo de producción del habla humana mediante el uso de un fuelle con bombeo de aire que imitaba el pulmón humano a través de diversas placas y compartimientos modulados que empleaban un teclado de 16 teclas similar al de un piano. La máquina podía formar cualquier palabra en varios idiomas europeos. Estos primeros sistemas eran mecánicos, no fue hasta final del siglo XIX cuando los científicos concluyeron que la voz se podía generar eléctricamente.

A principios del siglo XX J.Q. Stewart (Stewart, 1922) construyó una máquina que generaba sonidos vocálicos eléctricamente; y durante los años 30 se desarrollaron los primeros sistemas eléctricos que generaban todos los sonidos. El primero de ellos fue VOCODER, un analizador y sintetizador del habla desarrollado en los laboratorios Bell que se operaba mediante un teclado. Un operador humano experto podía seleccionar tanto una fuente periódica para los sonidos sonorantes como una fuente de ruido para los sonidos fricativos que podían ser modificados mediante un banco de filtros.

Al mismo tiempo aparecieron los primeros sistemas con capacidades básicas para el procesamiento de lenguaje natural.

Durante los años 40, se desarrollaron las primeras computadoras y algunos científicos prominentes como Allan Turing precisaron su potencial para el desarrollo de sistemas "inteligentes" y para medir la capacidad de inteligencia de las máquinas propuso el denominado "Test de Turing" (Turing, 1950), en el cuál un juez humano entabla una conversación en lenguaje natural con una máquina. Si el juez no era capaz de decir con fiabilidad si había hablado con una persona o una máquina, la máquina pasaba la prueba. Éste fue el punto de partida que fomentó las iniciativas de investigación que en los años 60 originaron los primeros agentes conversacionales. Por ejemplo, ELIZA de Weizenbaum (Weizenbaum, 1966),basado en la localización de palabras clave y el uso de plantillas predefinidas. Las plantillas transformaban la entrada del usuario en respuestas del sistema. Así, cuando el usuario escribía una frase como "Estoy X", ELIZA contestaba "¿Cuánto hace que estás X?" independientemente del significado de 'X'. Así, aunque su comportamiento fuera percibido como el de un ser humano por algunos usuarios novatos y pudo pasar el test de Turing, los primeros sistemas conversacionales tales como ELIZA no interpretaban semánticamente la entrada de los usuarios. Para tratar este reto, surgió durante los años 70 la investigación en lingüística computacional partiendo de los trabajos teóricos desarrollados desde los años 50 por Chomsky, Montague y Wood. Al mismo tiempo, aparecen los primeros sintetizadores del habla basados en reglas. En los años 70 también se desarrollan los primeros reconocedores continuos del habla basados en décadas de investigación sobre el habla discreta en la que los estímulos verbales se alternaban con pausas largas.

Beneficiándose de las mejoras incesantes en las áreas del reconocimiento del habla, el procesamiento del lenguaje natural y de la síntesis del habla, las primeras iniciativas de investigación relacionadas con los sistemas de diálogo oral surgen a principios de los años 80. El origen de este área de investigación está ligado a dos grandes proyectos: el programa DARPA Spoken Language Systems en los E.E.U.U. y Esprit SUNDIAL en Europa. El principal objetivo del proyecto DARPA fue el estudio y desarrollo de tecnologías relacionadas con el reconocimiento automático del habla y el procesamiento del lenguaje natural bajo el dominio de la reserva de vuelos vía telefónica, servicio que denominaron Air Travel Information Services (ATIS) (DARPA,

1992) (DARPA, 1994). El corpus de diálogos ATIS, que todavía utilizan los investigadores y desarrolladores de SDO, favoreció la aparición de otros proyectos como los llevados a cabo por AT&T. Por ejemplo, el proyecto AMICA (Pieraccini et al., 1997), donde se aplicaron diversos modelos estocásticos al desarrollo de SDO con iniciativa mixta. ATIS fue también el punto de partida para la investigación en el MIT y CMU, donde se han desarrollado algunos de los sistemas más importantes de la literatura.

Por otra parte, el proyecto SUNDIAL trataba con información sobre horarios de avión o tren en cuatro idiomas europeos. La investigación realizada en SUNDIAL fue el origen de numerosos proyectos financiados por la Comunidad Europea relativos principalmente al modelado del diálogo, por ejemplo, VERMOBIL (Bos et al., 1999), DISC (Bernsen y Dybkjaer, 1997) y ARISE (den Os et al., 1999). En el proyecto ARISE se desarrollaron simultáneamente seis sistemas: dos prototipos en italiano basados en la tecnología del CSELT (Castagneri et al., 1998) (Baggia et al., 2000), un prototipo francés desarrollado por el LIMSI (Lamel et al., 2000b) y dos prototipos en holandés y en francés basados en la tecnología de Philips. El proyecto DARPA ATIS ha sido considerado por algunos autores (Bangalore et al., 2006) como perteneciente a una generación previa de SDOs en comparación con SUNDIAL, pues al contrario que el segundo, el primero está restringido a un dominio cerrado del uso.

Entre los programas de investigación más importantes de los años 90 con capacidad multidominio, destaca DARPA Communicator. Este proyecto con financiación gubernamental tuvo como objetivo el desarrollo de tecnologías del habla novedosas que pudiesen emplear como entrada no sólo voz sino también otras modalidades. Los sistemas desarrollados en este programa por los participantes en E.E.U.U. y Europa eran capaces de interactuar con los usuarios en múltiples dominios, en los cuales el usuario y el sistema podían iniciar la conversación, cambiar de tema o interrumpirse. Por ejemplo, los investigadores de CMU desarrollaron el sistema Carnegie Mellon Communicator (Rudnicky et al., 1999), que permitía obtener información sobre itinerarios complejos que incluían tanto la reserva de vuelos y hoteles como el alquiler de coches. La arquitectura del sistema estaba basaba en diversos agentes especializados que permitían el desarrollo de módulos que funcionaban de forma totalmente independiente al encapsular la información dependiente de la tarea.

Las líneas de investigación más importantes durante los años 90 estuvieron relacionadas con la mejora de las tasas de éxito de los diversos módulos de los sistemas de diálogo. En primer lugar, con respecto al reconocimiento del habla, la mayor preocupación fue la robustez. Los autores investigaron la degradación de funcionamiento repentina que experimentaban los sistemas debido a cambios de menor importancia como variaciones en los micrófonos o en los canales de telecomunicación, e indicaron que la tecnología utilizada en ese tiempo no era capaz de ofrecer soluciones aceptables. Para superar estos problemas, los esfuerzos internacionales investigación se dirigieron principalmente hacia temáticas fundamentales como la robustez (Cole et al., 1995), estudios de cómo modelar las características espectrales (Holmes y Huckvale, 1994; Ostendorf et al., Sep 1996), mejora de lo modelos de coarticulación (Sun, 1997; Kirchhoff y Bilmes, 1999), modelado del "speech rate" (Pfau y Ruske, 1998; Morgan et al., 1997) y proporcionar métodos de reconocimiento independientes del locutor. El reconocimiento del habla robusto se definió como uno de los objetivos fundamentales en el dominio de DARPA ATIS (Stern et al., 1992). En Europa, la Acción COST 249[1] con un equipo investigador de 20 países europeos y desarrollada entre 1994 y 2000, investigó el reconocimiento continuo del habla recibida vía telefónica, tratando todas las temáticas mencionadas anteriormente, además de la selección de modelos acústicos, métodos de clasificación fonéticos y adaptación a las características del canal telefónico.

Adicionalmente, a finales de los 90, los sistemas tuvieron que mejorarse para hacer frente a diferencias entre el canal telefónico fijo y los teléfonos móviles, cada vez más populares. Los nuevos sistemas requerían la capacidad de trabajar con anchos de banda estrechos y bajas relaciones señal-ruido. Además, la incorporación de los teléfonos móviles implicó tener que ocuparse de una variedad cada vez mayor de entornos desde los que los usuarios podían interaccionar con los sistemas, con una robustez cada vez más exigente para poder gestionar la comunicación en ambientes muy ruidosos (Kacic, 1999). Por este motivo, uno de los estudios principales para el desarrollo del sistema SDR TREC-8 en el LIMSI trató el desarrollo de técnicas de reducción de ruido (Gauvain, 1999).

[1] http://www.elis.ugent.be/cost249/

Las líneas de investigación principales en el campo de NLU durante los años 90 estuvieron relacionadas con el trabajo con vocabularios más ricos, desplazándose así desde el reconocimiento de palabras aisladas al del habla espontánea. En este caso, las aproximaciones también se centraron en el estudio de métodos y algoritmos para la mejora de los análisis lingüísticos o intentar evitarlos utilizando técnicas de "word-spotting" más sofisticadas (Zue y Glass, 2000).

En relación con la gestión de diálogo, durante los años 90 se desarrollaron grandes esfuerzos hacia la consecución de diálogos menos restrictivos en los cuales los usuarios pudiesen llevar la iniciativa de la comunicación. Así, los autores consideraron el uso del "barge-in" no sólo desde el punto de vista del reconocimiento de habla, sino también desde la perspectiva de la interacción (Zue y Glass, 2000). Bangalore et al. (2006) describe tres generaciones de sistemas de diálogo oral en relación con, entre otras características, las iniciativas de gestión del diálogo y las capacidades de comprensión del lenguaje natural. En primer lugar, se describe una primera generación en la cual los SDOs utilizan iniciativas del sistema y la semántica está asociada directamente a la detección de palabras clave. La segunda generación comprende los sistemas de diálogo con iniciativa mixta en los cuales la comprensión de lenguaje natural se realiza mediante frames, lo que permitía a los usuarios hablar naturalmente dentro de un dominio específico. La tercera generación incluye SDOs que abordan múltiples tareas o dominios simultáneamente y que pueden mejorarse incluyendo capacidades multimodales y multimedia.

Una de las tendencias principales durante los años 90 estuvo relacionada con la definición de lenguajes estándar para el desarrollo de SDOs (Zue y Glass, 2000). De este modo a finales de 1999 el W3C Voice Browser working group presentó los primeros estudios de requisitos para navegadores web que supusieron la base de los lenguajes de etiquetas (p.ej. VoiceXML) para el desarrollo de sistemas de diálogo oral[2]. La modularización de los sistemas para obtener componentes más portables y reutilizables se llevó a cabo principalmente mediante el desarrollo de subdiálogos de uso frecuente. No fue hasta finales de los 90 cuando aparecieron las primeras arquitecturas para desarrollar componentes "plug-and-play". Por ejemplo, en 1998 apareció la arquitectura Galaxy, uno de los trabajos pioneros para el desarrollo de componentes de SDOs totalmente independientes (Seneff et al., 1998).

[2]http://www.w3.org/TR/voice-dialog-reqs/

Además, hubo una actitud positiva generalizada hacia la adopción de aproximaciones estocásticas para obtener métodos no supervisados que aumentasen las capacidades de los módulos de reconocimiento automático del habla y procesamiento del lenguaje natural (Glass, 1999). Así, existió una necesidad cada vez mayor por disponer de recursos lingüísticos y de corpus de datos compartidos con los cuales entrenar los algoritmos. Durante los años 90 surgen los primeros desarrollos de corpus y herramientas para la evaluación de sistemas proporcionando importantes recursos compartidos, como por ejemplo WordNet.

Durante la historia de los sistemas de diálogo oral, algunos expertos se han atrevido a prever cuáles serían las líneas futuras de investigación en el área (véase la figura 1.2). Estos objetivos se han desplazado gradualmente hacia metas cada vez más complejas. Tal y como se ha reflejado en los resultados de investigación comentados previamente, durante los años 90 las tendencias principales se centraron en aumentar la robustez de los diversos módulos del sistema (RAH, PLN, GD, GLN, CTH) (Cole et al., 1995; Kacic, 1999; Zue y Glass, 2000; Mangold, 2001). A partir de 2003 los expertos han propuesto objetivos de más alto nivel, tales como proveer al sistema de razonamiento avanzado, capacidad de resolución de problemas, facultad de adaptación, proactividad, inteligencia afectiva, multimodalidad y multilingüismo (Dale, 2003; Jokinen, 2003; Gao et al., 2005; Haas et al., 2005; Minker et al., 2006b,a). Como puede observarse, estos nuevos objetivos se refieren al sistema en su conjunto y representan tendencias importantes que se alcanzan en la práctica a través del trabajo común en diversas áreas y diversos componentes del sistema de diálogo. Así, para que un sistema sea multilingüe, tiene que poder reconocer las elocuciones de los usuarios en diversos idiomas, interpretar su contenido semántico usando conocimiento lingüístico de cada uno de ellos y generando lenguaje natural y voz sintetizada en todos estos idiomas. Así, en contraste con lo que sucedió en los años 90 cuando los objetivos se definieron para cada área (RAH, PLN, GD, GLN, CTH), las tendencias actuales de investigación se definen mediante grandes objetivos compartidos entre los diversos investigadores que trabajan en sus áreas respectivas hacia una meta común.

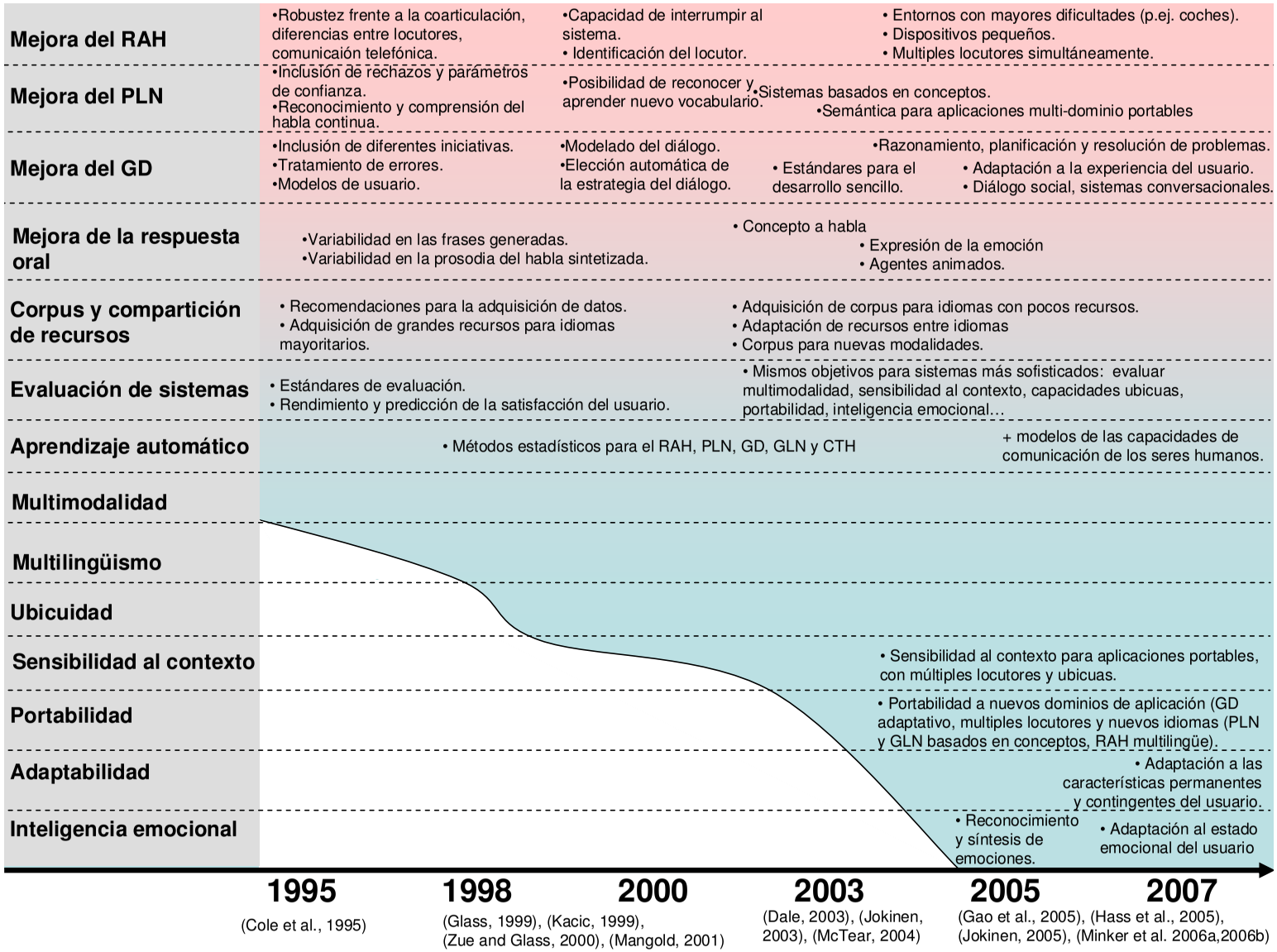


Figura 1.2. *Líneas de investigación descritas como futuro de los sistemas de diálogo oral*

De este modo, los expertos prevén sistemas de diálogo futuros inteligentes, adaptables, dinámicos, portables y multimodales. Todos estos conceptos no son mutuamente excluyentes, así la inteligencia del sistema está relacionada con su capacidad para adaptarse a nuevas situaciones y ésta a su vez le confiere mayor portabilidad para facilitar su uso en diversos entornos.

La proactividad es necesaria para que las computadoras pasen de ser consideradas herramientas y ser conviertan en verdaderas entidades conversacionales. Los sistemas proactivos tienen la capacidad de entablar una conversación con el usuario incluso cuando él no haya solicitado explícitamente la intervención de sistema. Se trata de un aspecto clave en el desarrollo de arquitecturas computacionales ubicuas en las cuales el sistema se encaja en el entorno del usuario de forma que éste no es consciente de estar interactuando con una máquina, sino que percibe que está interactuando con su entorno. En estas situaciones, la proactividad permite al sistema observar pasivamente el diálogo entre usuarios humanos y capturar el contexto conversacional relevante, que el sistema procesa para evaluar cuándo intervenir. Para alcanzar este objetivo, es necesario proveer los sistemas de capacidades de resolución de problemas y sensibilidad al contexto. Por ejemplo, en Schneider (2004) se describe un ayudante de compras dinámico integrado en carros de supermercado. El sistema observa las acciones y las intenciones del comprador para deducir los objetivos del usuario. Con esta información, la proactividad permite ofrecer ayuda adaptada al contexto actual como suministrar la información sobre productos cuando el usuario los sostiene durante mucho tiempo o comparar diversos productos cuando el usuario está decidiendo entre varios de ellos. Otros ejemplos de sistemas dinámicos se describen en (Baudoin et al., 2005; Kwon et al., 2005).

El interés por el desarrollo de sistemas con los cuales obtener una conversación tan natural y rica como entre seres humanos, fomentó la investigación sobre interfaces multimodales. En ellas, al contrario que con las interfaces tradicionales basadas en teclado y ratón o de sistemas de diálogo oral unimodales, existe flexibilidad en los modos de entrada y de salida (p.ej. gestos o expresiones faciales). La multimodalidad permite que los usuarios empleen diversas modalidades de la entrada así como obtener respuestas en diversos formatos, siendo especialmente importante para los usuarios con necesidades especiales para las cuales no son útiles las interfaces tradicionales. Los primeros sistemas de diálogo multimodales aparecieron a mediados de los

años noventa, combinando básicamente el habla con mapas gráficos (Cheyer y Julia, 1995; Oviatt et al., 1997) o con escritura (Waibel et al., 1997). Además, los movimientos oculares fueron unas de las primeras modalidades en ser estudiadas para ser empleadas para resolver ambigüedades acerca de a qué objetos se refiere el usuario cuando habla (Sarukkai y Hunter, 1997). La mayor parte de estos sistemas emplean la multimodalidad en la entrada o en la salida, en los últimos tiempos proyectos mayores se han dirigido hacia el desarrollo de la plena multimodalidad. Este es el caso del proyecto SMARTKOM, que proporciona lo que se ha denominado "multimodalidad simétrica total" en un sistema de diálogo con iniciativa mixta (Wahlster, 2006). Definen la simetría como la capacidad del sistema no sólo de entender y representar la entrada multimodal del usuario, sino también para generar una salida multimodal. La contribución principal del proyecto SMARTKOM no consiste únicamente en la integración o sincronización de las modalidades, sino también en tratar los fenómenos del diálogo que se asocian a la multimodalidad como la desambiguación mutua, la deixis multimodal y la resolución y generación de elipsis y anáfora multimodal.

La adaptabilidad puede referirse además a otros aspectos de las tecnologías del habla. En la interacción persona-ordenador, sobretodo en la multimodal, los usuarios disponen de diversas formas de comunicación. Los usuarios novatos y los usuarios experimentados podrían desear que el interfaz se comportara de forma totalmente diferente, como disponer de iniciativa por parte del sistema en lugar de iniciativa mixta. Un ejemplo de las ventajas de la adaptabilidad en el nivel de la interacción puede encontrarse en (Litman y Pan, 2002). Los usos multilingües son otro modelo de adaptación. Los reconocedores multilingües son capaces de reconocer simultáneamente varios idiomas compartiendo los modelos acústicos y/o los modelos de lenguaje. Los modelos acústicos multilingües consisten en un corpus de modelos acústicos dependientes del lenguaje para cada idioma, o una combinación de modelos acústicos independientes del idioma (Schultz y Kirchhoff, 2006). El desarrollo de un reconocedor del habla es una tarea costosa en términos de tiempo y esfuerzo, pues debe disponerse de una gran cantidad de datos de voz de centenares de locutores, anotados cuidadosamente para conseguir un conjunto representativo con el que entrenar los modelos acústicos. Para superar este problema, existe un interés creciente en la maduración de técnicas que permitan el desarrollo rápido de prototipos. La adaptación entre idio-

mas se presenta como una alternativa para compartir recursos en un idioma para el reconocimiento de otro, lo que es particularmente interesante para el desarrollo de reconocedores del habla para idiomas minoritarios. El estudio de técnicas para idiomas con pocos recursos es un área de investigación muy importante hoy en día, y muchos de los principales congresos en tecnologías del habla dedican explícitamente sesiones para estudiar esta temática, como por ejemplo Interspeech 2007 Special Session on Speech and language technology for less-resourced languages (2007) y LREC 2008 Workshop on Collaboration: interoperability between people in the creation of language resources for less-resourced languages (2008).

Tal y como describe Jokinen (2003), existen diversos niveles en los que el sistema puede adaptarse al usuario. El más simple es a través de perfiles personales en los cuales los usuarios disponen de opciones estáticas para personalizar la interacción (como decidir si desean que el sistema disponga de una voz masculina o femenina), funcionalidad que puede ampliarse clasificando a los usuarios en grupos de preferencias. Los sistemas pueden también adaptarse al entorno de los usuarios, como los sistemas proactivos ubicuos descritos anteriormente. Una aproximación más sofisticada consiste en la adaptación de los sistemas a diversos niveles de destreza del usuario (Haseel y Hagen, 2005), conocimiento del usuario (Forbes-Riley y Litman, 2004b) y necesidades especiales del mismo, como personas que emplean el sistema en un idioma que no es su lengua materna (Raux et al., 2003). Esta última temática está recibiendo mucha atención en cuanto al desarrollo de sistemas utilizables por personas con discapacidades (Heim et al., 2007), niños (Batliner et al., 2004) y personas mayores (Langner y Black, 2005). A pesar de su complejidad, estas características son hasta cierto punto aspectos estáticos; Jokinen identifica otro grado de adaptación en el cual el sistema no sólo se adapta al mensaje suministrado por el sistema, sino también a las intenciones del usuario y el estado de usuario. Siguiendo esta aproximación, la computación afectiva estudia cómo reconocer y adaptarse al estado emocional del usuario durante su conversación con el sistema.

Existe un interés cada vez mayor en el desarrollo de sistemas de diálogo oral que adapten dinámicamente su comportamiento al estado afectivo de los usuarios. Martinovski y Traum (2003) demostraron por medio de diálogos con usuarios utilizando un sistema de entrenamiento y un sistema de información vía telefónica, que se podrían evitar muchas interrupciones en

la comunicación persona-máquina si la máquina pudiese reconocer el estado emocional del usuario y tenerlo en cuenta para responderle. De hecho, experimentos anteriores (Prendinger et al., 2003) habían demostrado que la incorporación de un agente empático puede contribuir a una opinión más positiva acerca de la interacción. Polzin y Waibel (2000) presentan una primera aproximación para ajustar el comportamiento del diálogo al estado emocional del usuario, sugiriendo que se le ofrezca una realimentación más explícita si se detecta que puede frustrarse. Sin embargo, su aproximación define únicamente un conjunto de reglas de selección y no se basa en un marco general para el diálogo afectivo. Walker et al. (1997) examinan cómo factores sociales, tales como el estatus, influencian al contenido semántico, la forma sintáctica y la información acústica de las conversaciones. La inteligencia emocional incluye la capacidad de reconocer el estado emocional del usuario así como la capacidad de actuar sobre él apropiadamente. Actualmente, existe un conjunto de proyectos cuyo objetivo principal es dotar a los sistemas de diálogo con inteligencia emocional, algunos de los más recientes son MEGA (Camurri et al., 2004), NECA (Gebhard et al., 2004), VICTEC (Hall et al., 2005), NICE (Corradini et al., 2005), HUMAINE (Cowie y Schröder, 2005) y COMPANIONS (Wilks, 2006).

En cuanto a la portabilidad, ésta se trata actualmente desde perspectivas muy distintas, las tres principales son la independencia del dominio, del idioma y de la tecnología. Idealmente, los sistemas deberían poder trabajar en diferentes dominios de aplicación, o al menos ser fácilmente adaptables entre ellos. Los estudios actuales sobre la independencia del dominio se centran en cómo combinar las estructuras léxicas, sintácticas y semánticas de diversos contextos (Chambers y Allen, 2004) y cómo desarrollar gestores de diálogo que traten diferentes dominios (Mourão et al., 2004; Nguyen y Wobcke, 2006). En relación con la independencia del idioma, los sistemas multilingües (Schultz y Kirchhoff, 2006) son aquellos que pueden trabajar con varios idiomas, siendo así portables de dos formas principales: en primer lugar, los usuarios pueden suministrar la información en diversos idiomas, y en segundo lugar pueden recibir la respuesta también en idiomas diferentes, lo que es especialmente útil en sistemas de voz-a-voz, que pueden servir como intérpretes en tiempo real de modo que un usuario pueda hablar por teléfono y su interlocutor reciba la información traducida a su idioma.

Finalmente, la independencia de la tecnología hace frente a la posibilidad de utilizar sistemas de diálogo con diversas configuraciones de hardware. La potencia de los ordenadores personales continuará aumentando, con costes cada vez más bajos para los componentes de procesamiento y de memoria. Los sistemas que realizan incluso las aplicaciones más sofisticadas del habla se desplazarán desde arquitecturas centralizadas a configuraciones distribuidas para, de este modo, poder trabajar con diferentes tecnologías subyacentes. Para lograr este objetivo, se han estudiado arquitecturas estándar, la principal alternativa, dado que recoge los esfuerzos académicos y comerciales, es la propuesta por del MMI Working Group del W3C[3], cuyo objetivo es proporcionar un marco general y flexible que asegure la interoperabilidad entre los componentes de diferentes fabricantes o de diferentes tecnologías.

1.4 Objetivos de la tesis

El objetivo principal definido para la tesis fue contribuir a la adaptabilidad y la portabilidad de SDOs, que son las principales líneas de investigación actuales propuestas por los expertos en el área. El trabajo presentado en la tesis consiste en el desarrollo de modelos y métodos para gestionar diferentes temáticas de las descritas anteriormente, y evaluar experimentalmente su funcionamiento con sistemas de diálogo reales. No todos los aspectos descritos están en el objetivo de esta tesis, sin embargo el reto ha consistido en no centrarse solamente en un aspecto particular, sino en estudiar tres de los más desafiantes: reconocimiento de emociones, adaptación al idioma y evaluación de campo.

- Reconocimiento de emociones. El objetivo principal fue encontrar los factores decisivos para el reconocimiento de las emociones que influyen en el reconocimiento humano y automático. La mayor parte de las aproximaciones del estado del arte se basan en el uso de emociones simuladas para tener un control férreo sobre los corpus utilizados para entrenar los procedimientos automáticos. El objetivo principal definido para la tesis consistió en utilizar interacciones naturales de usuarios con un sistema de diálogo real. Éste es un objetivo particularmente desafiante, puesto que en este caso las emociones se producen de forma

[3]http://www.w3.org/TR/mmi-arch/

más sutil y la proporción de elocuciones con contenido emocional es muy baja en comparación con los casos neutros. La tesis se centra en estudiar el reconocimiento de emociones negativas que pueden originar el fracaso de la interacción con los SDOs.

- Adaptación al idioma. Uno de los principales requerimientos de la comunidad científica ha sido la necesidad de disponer de recursos comunes y de formas de poder aprovechar al máximo los disponibles, lo que es vital para la investigación centrada en idiomas o dialectos minoritarios. El objetivo para la tesis fue desarrollar una técnica que permitiría la adaptación rápida de un reconocedor del habla para poder reconocer un idioma diferente sin la necesidad de construir nuevos modelos acústicos. Para asegurar su portabilidad, el objetivo consistió en validarla no sólo entre idiomas con orígenes similares, sino también con idiomas muy diferentes.

- Evaluación de campo. Para poder desarrollar sistemas que se adapten a las necesidades y expectativas de usuarios es importante considerar sus opiniones sobre sus interacciones anteriores con el sistema. En la literatura, la mayoría de las evaluaciones se realizan con estudios de laboratorio y consideran por separado parámetros de la interacción y opiniones subjetivas sobre la calidad de la misma. El objetivo principal de la tesis fue descubrir relaciones significativas entre los parámetros de la interacción y las opiniones subjetivas mediante un estudio de campo, sobre interacciones no restringidas de los usuarios con un sistema real.

1.5 Estructura de la tesis

Tras esta breve descripción sobre los desafíos principales de la investigación para la creación de sistemas de diálogo adaptables y portables, las contribuciones de la tesis se describen en los capítulos 2 a 5. Cada uno de los capítulos se estructura como sigue: comienzan con una introducción que explica el objetivo de la investigación realizada; seguidamente se presenta un estado del arte detallado de la temática específica del capítulo y se comparan las contribuciones de la tesis con el trabajo disponible en la literatura. La tesis se apoya en una aproximación empírica donde todas las metodologías se evalúan con sistemas de diálogo reales. En cada uno de los capítulos hay una

sección dedicada a explicar la experimentación realizada, seguida por otras que presentan los resultados y las conclusiones experimentales extraídas de ella.

El capítulo 2 describe el sistema de diálogo oral UAH. El sistema se desarrolló cuidadosamente utilizando gestión dinámica del diálogo y diversas iniciativas de diálogo y estrategias de confirmación. El objetivo fue estudiar el efecto de cada una de ellas en las opiniones del usuario sobre sus interacciones con el sistema. En este capítulo, se describe la funcionalidad y la estructura del sistema UAH, describiéndose detalladamente cada uno de sus módulos poniendo un especial énfasis en sus funcionalidades innovadoras. El sistema está accesible al público vía telefónica y se han registrado y posteriormente anotado las llamadas de los usuarios. El corpus resultante se ha utilizado para la investigación sobre la evaluación de campo de los SDOs descrita en el capítulo 5, que describe detalladamente el procedimiento seguido para la anotación. El corpus UAH también se ha utilizado para la investigación sobre el reconocimiento de emociones en el capítulo 3. De este modo, UAH se utilizó como banco de pruebas de las diferentes metodologías propuestas en estas líneas.

El capítulo 3 presenta la investigación llevada a cabo para el reconocimiento automático de emociones, para la cual se ha realizado un estudio detallado sobre el impacto de considerar información contextual para la anotación de las mismas. En concreto, se propone la consideración de la historia del diálogo y del estilo neutro de los usuarios. Con respecto a la anotación humana, se ha dedicado especial atención al cálculo fiable del acuerdo entre anotadores. Respecto al aprendizaje automático, se ha desarrollado un nuevo método para incluir automáticamente ambas fuentes de información haciendo uso de nuevas técnicas para la normalización acústica y la anotación del contexto del diálogo. Las aproximaciones desarrolladas se han evaluado con un corpus de diálogos generado a partir de la interacción de aproximadamente 60 usuarios con el sistema UAH.

El capítulo 4 describe la investigación llevada a cabo sobre la adaptación entre idiomas. Se ha desarrollado una metodología para posibilitar que un sistema de reconocimiento del habla existente pueda utilizarse para el reconocimiento de otros idiomas diferentes. Esta metodología se ha evaluado con el sistema MyVoice, desarrollado en la Technical University of Liberec (República Checa). La aproximación se ha evaluado con tres idiomas: che-

co, español y eslovaco, para averiguar así si era posible utilizar el método propuesto no sólo con idiomas con orígenes similares (en este caso checo y eslovaco), sino también con idiomas que pertenecen a ramas muy diferentes de la familia de lenguas indoeuropeas (como el checo y español). El método propuesto consiste en cuatro pasos: 1) crear una correspondencia inicial entre los fonemas del idioma original y los del idioma objetivo, 2) crear un léxico y corresponder automáticamente la pronunciación de las palabras al conjunto original de fonemas, 3) ajustar los modelos para los fonemas que son únicos para cada idioma específico, y 4) recopilar datos en el nuevo idioma objetivo y realizar la adaptación al locutor. La investigación descrita en este capítulo se realizó en colaboración con investigadores de la Technical University of Liberec (República Checa), como resultado de tres meses de estancia investigadora en el Laboratory of Computer Speech Processing[4] bajo la supervisión del Prof. Jan Nouza.

El capítulo 5 muestra el trabajo llevado a cabo para la evaluación de campo del sistema de UAH. La evaluación de los sistemas de diálogo oral se ha realizado tradicionalmente mediante medidas instrumentales o medidas evaluadas por expertos (denominada generalmente evaluación "objetiva"), y las opiniones proporcionadas por los usuarios que han interactuado previamente con el sistema (denominada habitualmente evaluación "subjetiva"). Se han llevado a cabo diferentes trabajos para extraer relaciones entre ambos criterios de evaluación. En este capítulo se describen los resultados obtenidos mediante estudios que se realizaron a partir de las interacciones de usuarios reales con un SDO real, aspecto que no suele presentarse en la literatura.

La tesis termina con el capítulo 6, que presenta las conclusiones globales de la misma, una descripción detallada de las contribuciones principales, y un resumen de las pautas claves para el trabajo futuro. Finalmente, el Apéndice presenta las publicaciones desarrolladas a raíz de la investigación descrita en la tesis.

[4]http://itakura.kes.tul.cz/kes/indexe.html

Puisque c'est ell que j'ai écoutée se plaindre,
our se vanter, ou même quelquefois se taire.
Puisque c'est ma rose

Antoine de Saint-Exupery, Le Petit Prince

2 El sistema de diálogo oral UAH

2.1 Introdución

En el presente capítulo se describe el sistema de diálogo oral Universidad al Habla (UAH). En primer lugar, en la sección 2.2 hay una descripción detallada de la arquitectura del sistema describiendo sus módulos principales y los principios de diseño seguidos, incluyendo el acceso eficiente a bases de datos basado en el contexto de interacción y la gestión dinámica del diálogo. La sección 2.3 propone un enfoque novedoso para la creación de gramáticas de reconocimiento a partir de bases de datos que permite la reducción del tiempo empleado durante el proceso de reconocimiento, de esta forma se obtiene una respuesta del sistema rápida a las intervenciones del usuario y se incrementa la satisfacción de éste, al ser la interacción más fluida.

El desarrollo del sistema UAH se llevó a cabo por una parte con la finalidad de diseñar e implementar diferentes estrategias de gestión del diálogo en un sistema de diálogo real y por otra con el objetivo de emplear un sistema de diálogo real como banco de pruebas en el que evaluar las distintas metodologías propuestas en la tesis. El sistema ha estado a disposición del público en un número de teléfono local desde junio de 2005, de forma que los usuarios pueden interactuar con UAH para obtener información del Departamento de Lenguajes y Sistemas Informáticos. Todas las llamadas al sistema han sido grabadas y se han utilizado diversos criterios de evaluación que se almacenan en una base de datos, para los experimentos descritos en la tesis se ha empleado un corpus de 422 llamadas correspondiente a un año

de utilización del sistema UAH junto con la opinión personal de los usuarios que rellenaron cuestionarios subjetivos. La sección 2.4 describe la información extraída de las interacciones así como de encuestas de opinión, dicho corpus se ha empleado para evaluar las diferentes aproximaciones presentadas en el resto de la tesis.

2.2 Arquitectura modular

UAH es un sistema de diálogo oral desarrollado para dar acceso oral a la información académica del Departamento de Lenguajes y Sistemas Informáticos, así como otra información adicional de la Universidad de Granada y que se puso a disposición pública en junio de 2005. La figura 2.1 muestra la arquitectura modular de UAH. Como puede observarse, está compuesta de los cinco módulos típicos de los sistemas de diálogo actuales, que realizan el reconocimiento automático del habla, la gestión del diálogo, el acceso a bases de datos, el almacenamiento de datos y la generación de la respuesta oral. Además, existe un nuevo módulo denominado GAG (Generación Automática de Gramáticas) que permite la creación automática de reglas gramaticales para el reconocimiento del habla que se describe en la sección 2.3.

El gestor de diálogo decide la siguiente respuesta del sistema teniendo en cuenta los datos extraídos de las frases de usuarios y construye su respuesta empleando información extraída de bases de datos. Dicho comportamiento está codificado en documentos VoiceXML que se crean dinámicamente empleando PHP durante el transcurso del diálogo. UAH fue diseñado para dar información acerca de profesores, asignaturas, procesos de matrícula y estudios de tercer ciclo. Para llevar a cabo estas consultas se emplearon distintas iniciativas de diálogo con el fin de obtener medidas del grado en que la flexibilidad de la interacción afecta tanto al rendimiento del diálogo como a la satisfacción del usuario.

El gestor de diálogo del sistema UAH adapta las respuestas del sistema dinámicamente al contexto y estado del diálogo, por ejemplo para decidir la estrategia de confirmación a utilizar. El sistema utiliza confirmaciones explícitas para las acciones importantes o situaciones en que el proceso de reconocimiento es más difícil. En el resto de situaciones, el sistema emplea confirmaciones explícitas en el caso de que los valores de confianza no superen un umbral, y confirmaciones implícitas en el resto de las situaciones,

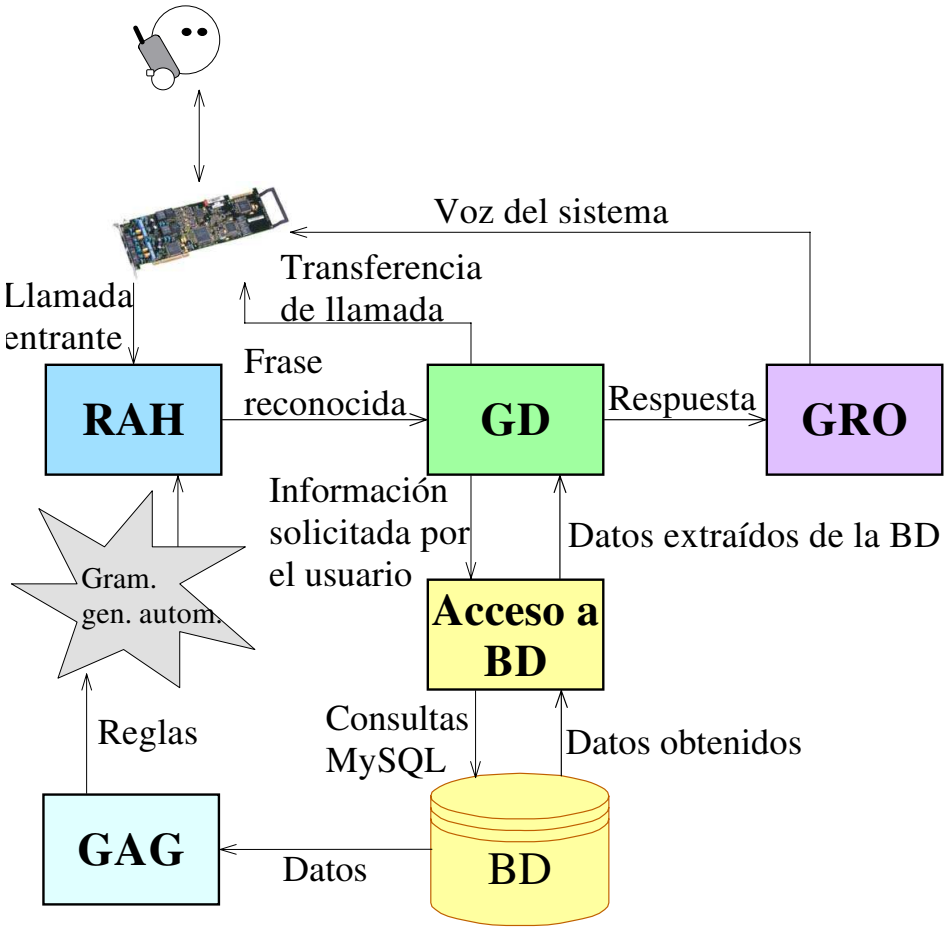


Figura 2.1. *Arquitectura modular del sistema Universidad Al Habla*

para ofrecer al usuario una interacción más natural. La generación de la respuesta oral del sistema se lleva a cabo en primer lugar en formato de texto instanciando patrones, así las frases se adaptan al contexto de interacción (p.ej. los mensajes de ayuda tienen en cuenta el tema tratado por el usuario y el sistema en un momento determinado). Una vez obtenida la respuesta en formato texto, se transforma a voz empleando un sintetizador de texto-a-voz comercial.

Con fines ilustrativos se muestra un diálogo ejemplo entre el sistema (S) y un usuario (U):

S1> Bienvenido al sistema UAH. ¿En qué puedo ayudarle?
U1> Necesito información acerca del proceso de matrícula en la universidad.
S2> ¿Qué desea saber acerca del proceso de matrícula?
U2> El plazo.
S3> El plazo de matrícula termina el 31 de agosto. ¿Necesita alguna otra información acerca del proceso de matrícula?
U4> No, gracias.
S5> Gracias por utilizar el sistema UAH. Que tenga un buen día.

2.2.1. Reconocimiento automático del habla

El módulo de reconocimiento automático del habla procesa cada frase del usuario proporcionada por una tarjeta telefónica Intel Dialogic D/41JCT-LS. Esta tarjeta gestiona la llamada del usuario proporcionando como resultado una señal acústica que es transformada en una secuencia de palabras en modo texto por el reconocedor. El proceso de reconocimiento se lleva a cabo empleando gramáticas que representan las frases válidas que el usuario puede mencionar. En UAH, estas gramáticas se crean de distintas maneras dependiendo de su vocabulario y el momento de creación. En total hay cuatro métodos de creación considerando las combinaciones de: i) vocabulario conocido/desconocido en tiempo de diseño, es decir, vocabulario que nunca cambia o vocabulario que se actualiza constantemente, como por ejemplo el contenido en una base de datos, ii) creación estática o dinámica de gramáticas.

Las gramáticas estáticas con vocabulario conocido a priori se crean durante la fase de diseño del sistema de diálogo y no varían su contenido con las distintas ejecuciones del mismo. Su contenido se conocía a priori y se diseñó cuidadosamente. Estas gramáticas se emplean en el sistema UAH para menús estáticos que presentan la información sobre la cual el usuario puede preguntar. El sistema UAH también emplea gramáticas prediseñadas para el reconocimiento de números y respuestas Booleanas del tipo sí/no/verdadero/falso. Las gramáticas estáticas con vocabulario desconocido se describen en la sección 2.3.

Las gramáticas dinámicas se emplean de dos formas: por una parte, pueden tener contenido conocido a priori, que se genera dinámicamente durante la ejecución del sistema sólo cuando realmente va a ser utilizado. Por otra parte, UAH también emplea gramáticas que se crean dinámicamente a partir de vocabulario extraído de bases de datos. Este vocabulario está en continuo cambio y, por tanto, no se conoce en la fase de diseño del sistema. Esta aproximación se emplea cuando las gramáticas contienen un vocabulario pequeño, puesto que crear gramáticas para vocabularios grandes en tiempo de ejecución introduciría un retardo en la interacción. En UAH este tipo de gramáticas se han empleado para desambiguaciones.

2.2.2. Gestión de diálogo

El gestor del diálogo decide la siguiente intervención del sistema de acuerdo con los datos extraídos de las entradas del usuario. La complejidad de este módulo depende de diversos factores como la complejidad de la interacción modelizada, el tipo de tarea a llevar a cabo mediante el diálogo, la flexibilidad deseada para el diálogo y el tipo de iniciativa implementada (dirigida por el usuario, dirigida por el sistema o mixta).

En el caso del sistema UAH, el dominio de aplicación consiste en la consulta de información en un entorno universitario. La iniciativa ha sido por otra parte deliberadamente dividida en distintos grados de flexibilidad entre los diferentes escenarios del sistema, de esta forma se pudieron extraer medidas del impacto de la flexibilidad del diálogo en la calidad observada y los parámetros de interacción (capítulo 5). En concreto, para el acceso a la información acerca de profesores y asignaturas se definió una iniciativa dirigida por el sistema, mientras que los diálogos que aportaban información sobre matrículas y estudios de tercer ciclo tenían iniciativa mixta.

Aunque no hay consenso para definir las tareas que debe realizar un gestor de diálogo, una de las aproximaciones más ampliamente aceptadas es la presentada por Traum y Larsson (2003). Los autores consideran que el gestor de diálogo debe actualizar el contexto del diálogo para obtener las interpretaciones semánticas correctas de las intervenciones de los usuarios. Además, el gestor del diálogo debe realizar sus tareas en un dominio específico (en este caso, de información académica) y decidir qué información proporcionar al usuario, cuándo y cómo expresarla.

El gestor de diálogo de UAH adapta las respuestas del sistema dinámicamente al contexto y el estado del diálogo, expresando de forma diferente algunas frases para mejorar la naturalidad de la interacción. Por ejemplo, los mensajes de ayuda del sistema tienen en cuenta el tema que el usuario y el sistema están tratando en un momento particular.

El contexto se emplea asimismo para decidir qué estrategia de confirmación usar. El sistema emplea confirmaciones explícitas para acciones importantes o situaciones en que el proceso de reconocimiento del habla es más difícil. Así, antes de transferir una llamada a un profesor el sistema confirma explícitamente su nombre. En el resto de situaciones el sistema emplea confirmaciones implícitas para permitir al usuario más flexibilidad y naturalidad en su interacción con el sistema (Bernsen et al., 1994).

2.2.3. Acceso a base de datos

En la literatura podemos encontrar múltiples referencias a la importancia de separar el acceso y consulta a bases de datos del resto de las tareas llevadas a cabo por el sistema de diálogo. Siguiendo esta filosofía, en el proyecto GEMINI (Hamerich et al., 2004) se construyó un asistente para conectar a las bases de datos de forma que los usuarios pudieran crear sistemas de diálogo de forma semi-automática independientemente de las características de las bases de datos empleadas. En UAH obtuvo la dicotomía deseada entre la gestión de diálogo y acceso a la información mediante la creación de un módulo de acceso a bases de datos. El gestor de diálogo proporciona a este módulo la información que el usuario necesita saber, a partir de ella el módulo de acceso construye la consulta y extrae los datos correspondientes de la base de datos.

Una vez que se ha extraído la información, el módulo de acceso ejecuta un programa PHP que valida los datos. Además, comprueba que no hay datos repetidos en el resultado antes de devolver la respuesta al gestor de diálogo. Finalmente, el gestor de diálogo decide cómo comunicar los datos al usuario. Sin embargo, hay situaciones en que el módulo de acceso de UAH no sólo recibe la información extraída (p.ej. el teléfono de un profesor apellidado "García"), sino también restricciones como el sexo. Por ejemplo, si el usuario pregunta por el número de teléfono de la Sra. García, entonces sólo se extrae datos acerca de profesoras (no profesores) con dicho apellido.

El sistema UAH hace dos tipos de consultas a bases de datos: explícitas e implícitas. En el primer caso, la consulta se lleva a cabo por iniciativa del usuario. De este modo, si el usuario pregunta el número del profesor "José García", se realizan dos consultas: la primera comprueba el número de registros en la base de datos que corresponden a profesores con dicho nombre. Si el número es cero, el sistema informa al usuario que no se encontraron profesores con dicho nombre. Si el número es mayor de uno, el sistema pide al usuario información adicional que permita seleccionar al correcto. Una vez que se ha determinado el nombre, se realiza otra consulta para extraer el número de teléfono. La primera consulta está implícita porque es iniciada por el sistema sin que haya una petición explícita del usuario. Por el contrario, la segunda es explícita porque se corresponde con la petición del usuario.

Adicionalmente, la complejidad de las consultas en UAH varía con la flexibilidad del estado de diálogo actual. De hecho, las consultas pueden variar desde un simple "select FIELD from TABLE" a una selección de datos compleja basada en la concordancia de cadenas completa o parcial. Por ejemplo, cuando el usuario pronuncia el nombre de un profesor, la información aportada puede ser una combinación de nombre y apellidos hasta siete posibilidades, incluyendo también nombres incompletos (p.ej. "José" en lugar de "José Luis").

El sistema extrae información consultando una base de datos que contiene información pública acerca de la Universidad de Granada (p.ej. número de fax y correo electrónico de profesores, pero no sus datos personales). La base de datos se diseñó para almacenar los datos que se suelen aportar en las páginas web de los departamentos. El sistema UAH podría trabajar con cualquier base de datos que contuviera muchos tipos de información, creando vistas si es necesario para mantener la privacidad de datos económicos y personales.

2.2.4. Generación de respuesta oral

La generación de la respuesta del sistema se lleva a cabo usando patrones que podemos clasificar en las siguientes categorías: profesor, tercer ciclo, asignatura, matrícula, información adicional, confirmación, saludos y ayuda. Una vez que se obtiene la respuesta en modo texto, se transforma a voz empleando el sistema texto-a-voz comercial de Verbio.

Dentro de cada categoría de respuesta hay un patrón específico para cada tipo de dato que el sistema puede ofrecer al usuario. Estos patrones están compuestos de varios segmentos de información que pueden ser seleccionados o ignorados dinámicamente según los datos que haya que aportar. De este modo, el patrón empleado para informar acerca de la localización del despacho de un profesor se compone de tres segmentos principales: número del despacho, nombre de la facultad donde está y piso dentro de la facultad. Si uno de estos datos no está disponible en el estado del diálogo actual, no se incluye en el texto resultante. Así, si la planta no está disponible la respuesta del sistema podría ser "despacho número 3 en la Facultad de Ciencias Empresariales").

Además, la estructura morfológica de los patrones puede variar dependiendo de la información a aportar, puesto que debe tenerse en cuenta el género (p.ej. "Sr. Juan López" o "Sra. Clara López") y el número (p.ej. "La asignatura programación paralela tiene dos créditos" o "La asignatura programación paralela tiene un crédito"). El sistema también emplea adaptación dinámica de nexos (p.ej. "Hay dos profesores llamados García: Pedro García y María García" o "Hay tres profesores llamados García: Pedro García, María García y Javier García").

Finalmente, la última etapa en la creación de la respuesta del sistema en modo texto consiste en la adaptación de varias palabras especiales y símbolos, así como la inclusión de etiquetas de prosodia. La adaptación de palabras especiales transforma la información contenida en la base de datos a una forma susceptible de ser leída por el sistema, mientras que la inclusión de etiquetas permite un mejor entendimiento de la información sintetizada por el sistema (p.ej. es mejor deletrear una URL que leerla como si fuese una palabra).

2.3 Generación automática de gramáticas

La mayoría de los entornos de desarrollo de sistemas de diálogo comerciales proporcionan herramientas para la creación y prueba de gramáticas basadas en reglas. En estas herramientas, la creación de gramáticas se lleva a cabo de forma estática; es decir, en tiempo de diseño, con anterioridad a la puesta en servicio del sistema. Aunque ésta es la forma más sencilla de crear gramáticas, la creación estática puede provocar inconsistencias entre

los contenidos de la base de datos y el vocabulario incluido en las gramáticas cuando el vocabulario es desconocido a priori o no se mantiene inalterado.

Para solventar esta desventaja, existen diversas técnicas de creación dinámica de gramáticas. Una primera técnica las construye en tiempo de ejecución (Truillet et al., 2004). Esta técnica ofrece flexibilidad puesto que las gramáticas siempre están actualizadas con los últimos cambios en las bases de datos. Sin embargo, este método puede implicar una carga computacional muy elevada, que introduce un retraso que, en sistemas con bases de datos muy grandes puede ser excesivo, provocando que el sistema sea considerado "lento" por los usuarios.

Una segunda técnica crea las gramáticas al principio de cada interacción, antes del procesamiento del habla. Estas gramáticas siempre se actualizan con independencia de los cambios en la base de datos (Schalkwyck y Story, 2003). Este método no implica un incremento en el tiempo de reconocimiento. Por tanto, el incremento en el tiempo de espera es menor desde el punto de vista del usuario pero sigue habiendo un retraso en el inicio del sistema que incrementa el tamaño de la base de datos.

En ambos casos, la idoneidad de la técnica depende del tamaño del vocabulario. Como se ha mostrado anteriormente (Nielsen, 1994) existen tres tiempos de ejecución que deben ser considerados. En primer lugar, el límite a partir del cual el usuario considera que la interacción es en tiempo real, que es de 0,1 segundos. En segundo lugar, de 0,1 a 1,0 segundos, que es el intervalo en el cual el usuario nota un retraso pero su flujo de pensamiento no se ve interrumpido. Y en tercer lugar, desde 1,0 a 10,0 segundos, cuando el usuario todavía presta atención al sistema. Para retrasos mayores de 10 segundos es necesario dar algún tipo de respuesta al usuario con un mensaje del sistema o reproduciendo música (Cerrato, 2002).

El tiempo de ejecución necesario para construir la regla de una gramática a partir de una columna específica de una base de datos se midió respecto al número de palabras extraídas. Como se muestra en la figura 2.2, el límite de 0,1 segundos se alcanza con 10.000 palabras, el de 1 segundo con 100.000 y el máximo (10 segundos) se alcanza con vocabularios alrededor de 300.000 palabras. De esta manera, para un vocabulario de un millón de palabras, hay un retraso de 30 segundos. Por tanto, para vocabularios grandes (mayores de 300.000 palabras) el retardo introducido es apreciado por el usuario y es necesario establecer alguna técnica de retroalimentación. Sin embargo, dichas

técnicas normalmente causan una mala impresión a los usuarios, que en numerosas ocasiones las encuentran irritantes (Mäkelä et al., 2001). Además, es vital alcanzar un tiempo de respuesta mínimo para minimizar los costes de la interacción, entendiendo por tiempo de respuesta el tiempo transcurrido desde que el usuario termina de hablar y se presenta la respuesta del sistema (Möller, 2005)). Este hecho unido a la maximización del éxito de la tarea, es esencial para asegurar la satisfacción del usuario y se suele emplear como parámetro clave para la evaluación de sistemas de diálogo, tal y como propone el paradigma PARADISE (Walker et al., 2000a).

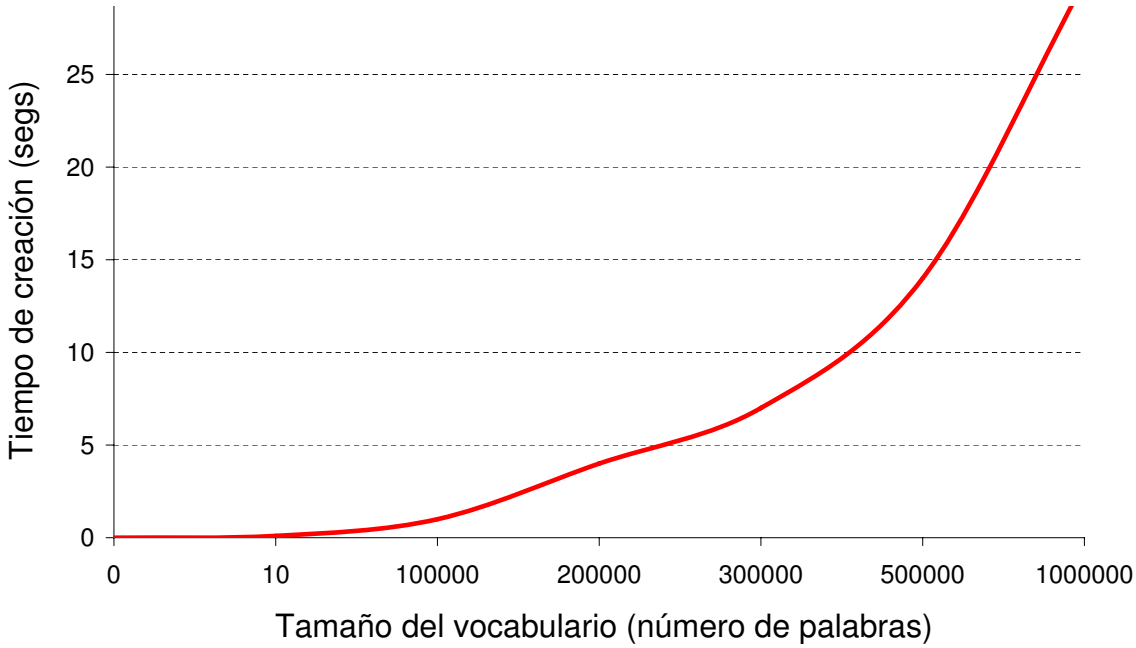


Figura 2.2. *Latencia de la creación dinámica de reglas gramaticales*

Para evitar los problemas comentados anteriormente, se propone una nueva técnica denominada CGBD (Creación de Gramáticas Basada en Disparadores) para optimizar así el proceso de creación de gramáticas para RAH basadas en reglas. La idea principal se basa en actualizar el vocabulario de dichas gramáticas de forma automática conforme se producen cambios en las bases de datos utilizadas, empleando para ello un mecanismo de disparadores. De esta forma, no hay retardo de creación en tiempo de ejecución y el usuario siente que está interactuando con un sistema rápido.

La mayor ventaja de esta técnica en comparación con otras que encontramos en la literatura (McTear, 2004) es que es general y, por tanto,

adecuada para crear gramáticas que puedan emplearse en cualquier sistema de diálogo que extraiga vocabulario de bases de datos. El proceso de creación de dichas gramáticas se lleva a cabo en tres pasos: i) extracción de información de las bases de datos, ii) construcción de las reglas gramaticales a partir de los datos extraídos disponiendo los datos en formato JSGF (Java Speech Grammar Format) o ABNF (Augmented Backus-Naur Form), iii) actualización de las gramáticas con los cambios efectuados en las bases de datos. Para asegurar que las gramáticas están actualizadas permanentemente incluso cuando cambia el contenido de las bases de datos, nuestra técnica emplea un mecanismo basado en disparadores de bases de datos. Éstos se disparan cuando se producen cambios en los campos de las bases de datos a partir de los cuales se extraen las palabras para las reglas gramaticales.

El grupo de trabajo W3C Voice Browser especifica en (Brown, 1999) que las reglas deben ser redefinibles en tiempo de ejecución. Para ello, muestran diversos mecanismos. Entre ellos, subrayamos la división del espacio de reglas en estáticas y dinámicas. Empleando la técnica CGBD, la estructura de frase se crea estáticamente conteniendo referencias a reglas dinámicas, que pueden almacenarse en el mismo fichero o en uno externo.

La técnica CGBD fue implementada en la herramienta GAG (Generación Automática de Gramáticas) empleando PHP, HTML, JavaScript y PostGreSQL. La herramienta se emplea como un módulo adicional del sistema de diálogo oral UAH. Dispone de una interfaz sencilla para permitir al diseñador del sistema escoger los campos de la base de datos que se emplearán para extraer el vocabulario. El proceso para crear reglas empleando esta herramienta se divide en tres etapas. En primer lugar, la herramienta GAG solicita al usuario el nombre de la base de datos, el nombre del ordenador servidor, el nombre de usuario y la contraseña para acceder a la base de datos donde se encuentra almacenada la información. También solicita el nombre de la regla de gramática y el tipo de gramática a crear (JSGF o ABNF). En segundo lugar, se visualizan los campos de las tablas de la base de datos en un menú desplegable. Los campos aparecen en el menú en el mismo orden en que se encuentran en la tabla correspondiente, pero el diseñador puede seleccionarlos en cualquier orden para crear la regla gramatical. Cuando se selecciona un campo, un número aparece automáticamente junto al campo indicando el orden de selección. Por ejemplo, en una aplicación académica el diseñador del sistema podría seleccionar el nombre de profesor y segui-

damente su apellido obteniendo la regla: <profesor>= ("Sergio García"|...| "Miguel Moreno"). Alternativamente, podría seleccionar ambos casos en orden inverso, es decir, primero el apellido y después el nombre, en cuyo caso la regla obtenida sería: <profesor>= ("García Sergio" |...| "Moreno Miguel"). Por ultimo, después de la selección de campos el diseñador debe introducir el nombre del archivo donde desea almacenar la regla de la gramática. Puede elegir, además, entre crear la regla como parte de una gramática existente en su mismo fichero, o bien, crear un único fichero con una nueva gramática cuya única regla sea esa.

GAG implementa el mecanismo de disparadores CGBD para mantener las reglas gramaticales actualizadas con los últimos cambios realizados en las bases de datos. Para ello, al final de la creación de cada gramática el diseñador del sistema indica si desea que se actualice automáticamente el vocabulario utilizando la regla que muestra los cambios en la base de datos. En caso afirmativo, se crean dinámicamente disparadores que se activan cuando los valores correspondientes se actualizan, borran o insertan. Así, si se crea una regla para los nombres y apellidos de profesores, el disparador incluirá "María" como una nueva palabra en la regla si se introduce en el campo "Nombre" de la tabla "Profesor". De idéntica forma, borra la palabra "María" de la regla si se borra de la tabla, y cambia "María" por "Mónica" si el campo se actualiza con este nuevo valor.

Como los disparadores indican los valores antiguos y nuevos del campo de la tabla, sólo se realizan los cambios pertinentes en las reglas de las diversas gramáticas en lugar de generarlas por completo de nuevo cada vez que ocurra cualquier cambio, lo que incluiría un retardo equivalente al de crearlas dinámicamente. El proceso completo de actualización automática se ilustra en la figura 2.3, donde se esquematiza una arquitectura distribuida en la que el RAH se realiza en un ordenador que dispone de una tarjeta de interfaz telefónica, mientras que el gestor de diálogo y la herramienta GAG se ejecutan en el ordenador servidor. Las bases de datos se almacenan en un equipo diferente.

En los experimentos realizados comparamos dos aproximaciones distintas para la creación de gramáticas para RAH basadas en reglas a partir de datos extraídos de bases de datos: creación dinámica (método tradicional) y creación usando la técnica propuesta (CGBD). Para realizar una comparativa lo más ilustrativa posible de ambas técnicas, se emplearon vocabularios

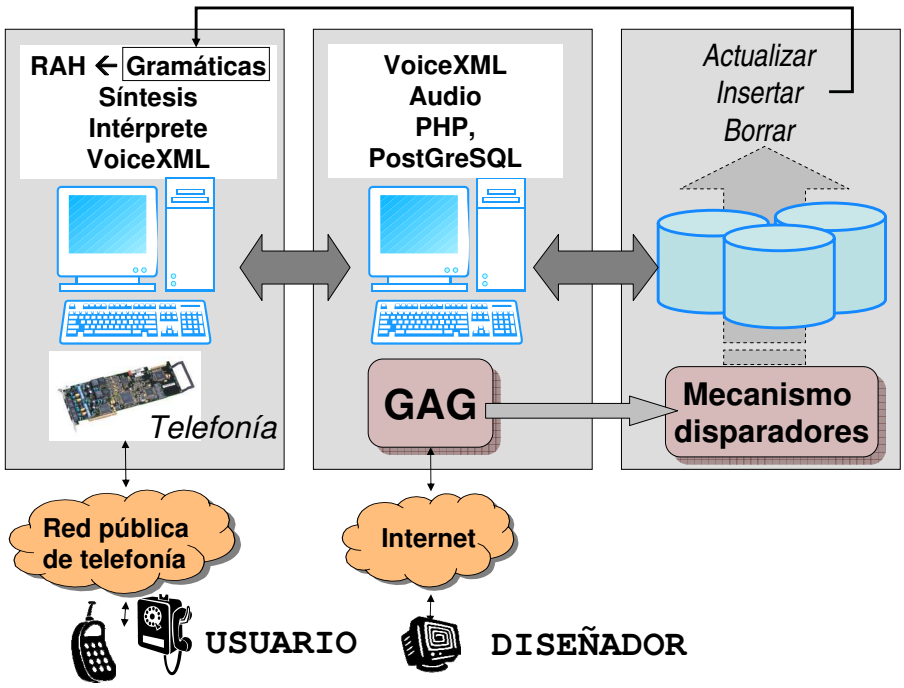


Figura 2.3. *Reglas gramáticales automáticas actualizadas con la herramienta GAG*

que contenían desde 1 a 10^6 palabras (como los mostrados en la figura 2.2). Sin embargo, el tamaño del vocabulario gestionado en UAH es inferior a las 4.000 palabras, con lo que en muchas ocasiones se tuvieron que emplear palabras de relleno generadas automáticamente en la base de datos desde la que se extrae el vocabulario para generar las reglas gramaticales de RAH. Utilizando configuración anterior, se solicitó a 30 usuarios que evaluaran el sistema dos veces.

En primer lugar con la técnica de generación dinámica, pidiendo que expresaran su grado de satisfacción (en una escala de 1 a 5 puntos) acerca de la velocidad de interacción con el sistema; y en segundo lugar, empleando la técnica CGBD, que eligieran entre ésta y la primera estrategia utilizada. Los resultados de los experimentos anteriores son los mostrados en la figura 2.4, donde se aprecia que el grado de satisfacción de los usuarios respecto a la creación dinámica de las gramáticas decae conforme aumenta el tamaño del vocabulario. En cambio, la tendencia es opuesta cuando se usa la técnica propuesta.

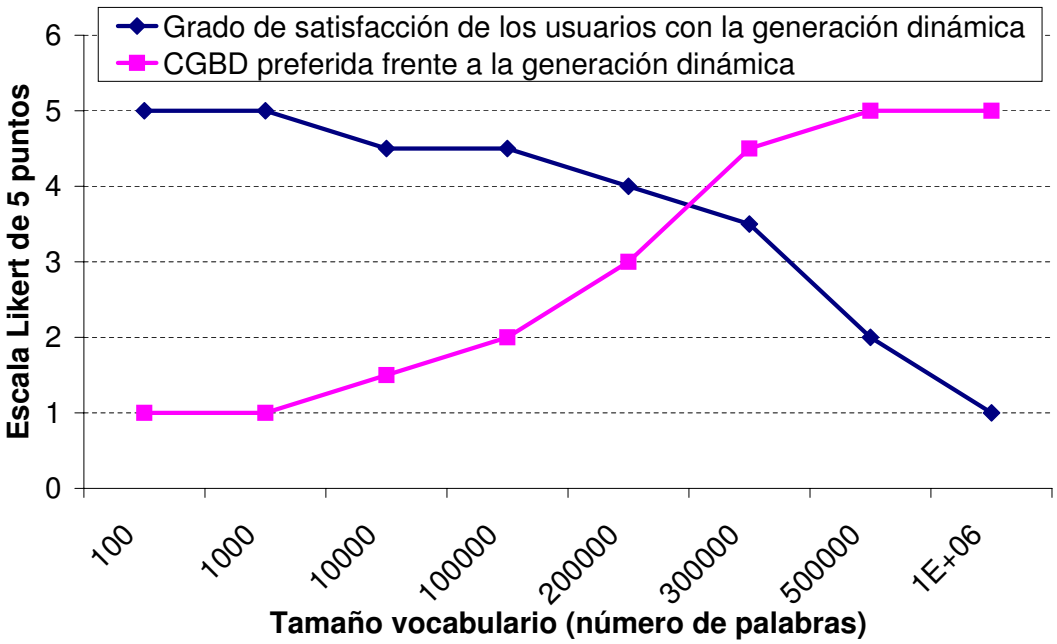


Figura 2.4. *Técnica CGBD versus la creación dinámica de reglas gramaticales en cuanto a la satisfacción del usuario*

Cabe destacar claramente la correspondencia entre el grado de satisfacción de los usuarios y las restricciones de tiempo comentadas. Para vocabularios menores de 10.000 palabras (que conllevan un tiempo de generación con la herramienta GAG menor de 0,1 segundos) los usuarios no son conscientes del retardo del sistema cuando se usa la técnica de generación dinámica. También puede observarse que existe un punto en el que ambas líneas se cruzan: cuando el vocabulario tiene un tamaño de unas 300.000 palabras. La figura muestra que a partir de este punto, el tamaño del vocabulario es lo suficientemente grande como para hacer que la técnica propuesta sea claramente preferible a la creación dinámica, dado el retardo causado por ésta última.

Además, respecto al empleo de la herramienta GAG, la percepción de los usuarios varía según la iniciativa de gestión de diálogo empleada. Aunque el 71 % de los usuarios encuentra la velocidad de interacción adecuada o rápida, el 20 % la considera lenta (muy lenta sólo el 3 %). Esto ocurre debido a que se empleó iniciativa dirigida por el sistema para esta experimentación. La figura 2.5 muestra que, en general, la satisfacción del usuario con la interacción es en el 78 % de los casos positiva.

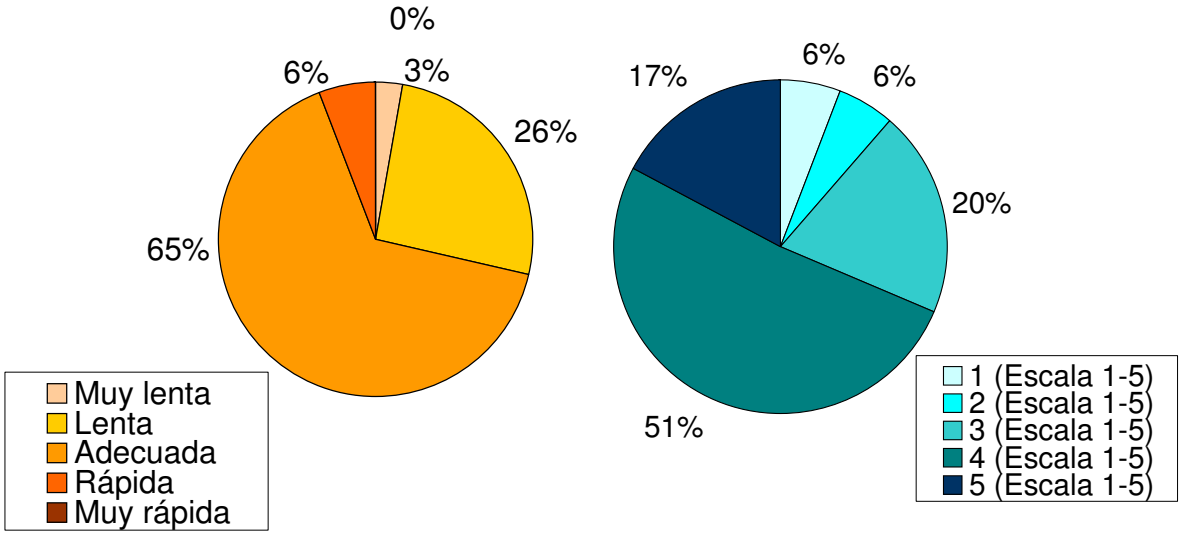


Figura 2.5. *Velocidad de la interacción percibida (izquierda) y satisfacción del usuario (derecha) utilizando UAH*

Se encontró además que, cuando las medidas de evaluación objetivas indicaban que el retraso en el sistema no era significativo, algunos usuarios opinaban lo contrario. Por tanto, puede que los usuarios consideren que una interacción es lenta simplemente por el tipo de iniciativa de gestión de diálogo (p.ej. por ser dirigida por el sistema), aunque el tiempo utilizado para la generación de la respuesta haya sido pequeño (ver capítulo 5 para una discusión ampliada). Este hecho se explica porque los usuarios sienten que la interacción podría ir más rápida si tuvieran la oportunidad de proporcionar directamente toda la información de sus peticiones.

2.4 El corpus oral UAH

El corpus empleado para los experimentos descritos en la presente tesis se compone de 85 diálogos de unos 60 usuarios diferentes que interactuaron con el sistema UAH. Contiene 422 turnos de usuario, con una media de 5 turnos por diálogo. En total comprende 150 minutos de grabación. El corpus se etiquetó de forma semi-automática empleando criterios estándares (de facto) para la evaluación y obteniéndose la opinión de los usuarios mediante cuestionarios, tal y como se describe en el capítulo 5. Posteriormente,

cada intervención de usuario se anotó con un estado emocional por nueve anotadores no expertos tal y como se describe en el capítulo 3. El tamaño del corpus es similar a otros corpus emocionales de la literatura como los empleados por Forbes-Riley y Litman (2004a) (10 dialogos, 453 turnos) o Morrison et al. (2007) (391 turnos de usuario).

2.5 Conclusiones

UAH es un sistema de diálogo desarrollado utilizando el estándar VoiceXML. Utiliza gestión dinámica del diálogo e introduce algunas técnicas novedosas que hacen la interacción más fluida. Se desarrolló para cubrir distintas iniciativas de interacción y estrategias de confirmación con el fin de evaluar su comportamiento. Cabe destacar la introducción del módulo GAG, que proporciona una nueva forma de creación de gramáticas de reconocimiento del habla sin introducir retardos y manteniendo el vocabulario actualizado. El sistema ha estado en funcionamiento desde junio de 2005, cuando se puso a disposición pública. Desde entonces se graban todos los diálogos de los usuarios y se procesaron semi-automáticamente los principales criterios de evaluación objetiva, además se obtienen criterios de evaluación subjetiva de encuestas de opinión realizadas tras la interacción con el sistema por algunos usuarios. A partir de esta información se ha construido un corpus oral que es la base en la que se sustenta la investigación descrita en los siguientes capítulos de la tesis.

Innegable, señor.
Es indisimulable.
¿Está usted aburrido?
Me parece que está usted aburrido.
Dígame, ¿adónde va tan aburrido?

Rafael Alberti, El aburrimiento

3

Reconocimiento de emociones no actuadas

3.1 Introducción

Uno de los principales objetivos de investigación en sistemas de diálogo es alcanzar un nivel de comunicación entre los seres humanos y las máquinas parecido al que existe entre las personas. Esto eliminaría la necesidad de emplear teclado y ratón en favor de formas de interacción más intuitivas como el lenguaje natural, conduciendo a un nuevo paradigma en que las tecnologías puedan ser accedidas por usuarios no expertos o discapacitados.

Sin embargo, la interacción hombre-máquina, aún siendo multimodal, todavía no es comparable al diálogo humano. Esto se debe entre otras cosas a que la interacción humana involucra no sólo el intercambio de contenido explícito, sino también de información implícita acerca del estado emocional del interlocutor. Los sistemas que hacen uso de este tipo de información se dice que incorporan "computación afectiva" (affective computing) o "inteligencia emocional" ("emotion intelligence"), campos que engloban el reconocimiento, interpretación, gestión y generación de emociones.

Debido a sus beneficios y a su gran variedad de aplicaciones, la computación afectiva ha emergido como una línea de investigación puntera en el campo de la interacción persona-ordenador, habiendo aparecido numerosos proyectos internacionales e interdisciplinares asociados. Algunos de los últimos son MEGA (Camurri et al., 2004), NECA (Gebhard et al., 2004), VICTEC (Hall et al., 2005), NICE (Corradini et al., 2005), HUMAINE (Cowie y Schröder, 2005) y COMPANIONS (Wilks, 2006), por mencionar algunos.

La anotación precisa es un primer paso para una detección y gestión optimizada de las emociones, tareas que son muy importantes para evitar problemas significativos en la comunicación, como falta de entendimiento o frustración del usuario, que pueden llevar al fracaso del diálogo. A pesar de sus beneficios, la anotación de emociones en los sistemas de diálogo oral encuentra restricciones como resultado de ciertos problemas importantes. En primer lugar, como se muestra en la sección 3.3.2, el porcentaje de discurso neutro frente al emotivo suele estar muy desequilibrado (Forbes-Riley y Litman, 2004a; Morrison et al., 2007). En segundo lugar, toda la información debe ser recogida a través de la modalidad oral y en algunos sistemas donde el diálogo es menos flexible, la duración de las intervenciones del usuario puede ser insuficiente como para permitir la utilización de otras fuentes de conocimiento como información lingüística.

Para resolver estos problemas, en la presente tesis se propone el uso de información contextual para la anotación de emociones de usuarios en sistemas de diálogo oral. El principal interés es reconocer la emociones negativas puesto que algunos autores, como Riccardi y Hakkani-Tür (2005), han mostrado que una vez que un usuario está en un estado emocional negativo, es difícil sacarlo de él. Además, estas malas experiencias pueden desanimar a los usuarios a emplear el sistema de nuevo. En concreto, se han tenido en cuenta tres emociones negativas. La primera es *duda*, que es útil para identificar cuándo el usuario no está seguro acerca de lo que hacer o decir a continuación. La segunda y tercera emoción son *enfado* y *aburrimiento* respectivamente, dos estados emocionales negativos que debe ser reconocidos antes de que el usuario se frustre irreversiblemente durante su interacción con el sistema. En el espacio de activación-evaluación (Russell, 1980; Scherer, 2005), *enfado* se corresponde con una emoción negativa activa, mientras que *aburrimiento* y *duda* son emociones negativas pasivas.

En este capítulo se presentan distintos enfoques para incluir información emocional tanto en la anotación humana de emociones (sección 3.3), como en la clasificación basada en aprendizaje automático (sección 3.4). En la anotación humana, anotadores no expertos reciben información contextual dándoles las intervenciones del usuario a anotar junto con los diálogos completos en los que se produjeron. De esta forma, los anotadores reciben información sobre la forma de hablar del usuario y de cómo había sido el desarrollo del diálogo hasta el momento de su intervención. Para la clasificación

automática, se introduce un método novedoso que mejora la clasificación de emociones negativas con información contextual generada de forma automática. En primer lugar, calcula el estilo de habla neutra del usuario, que se emplea para clasificar emociones en las categorías *enfado* y *dudaOaburrimiento* y a continuación introduce el contexto del diálogo para distinguir entre las categorías *aburrimiento* y *duda*. Una de las principales ventajas del método propuesto es que no requiere información adicional anotada manualmente. Por tanto, permite la integración automática de información contextual en reconocedores de emociones para sistemas de diálogo oral.

Para evaluar los beneficios de las propuestas, se han desarrollado diversos experimentos sobre un corpus de emociones reales extraído de la interacción de alrededor de 60 usuarios diferentes con el sistema de diálogo UAH (capítulo 2). El objetivo era demostrar que la información contextual propuesta influye la anotación humana y el reconocimiento máquina, y que se pueden obtener mejores resultados cuando se incluye información contextual empleando los métodos desarrollados en la tesis, en comparación con el reconocimiento basado en las características acústicas tradicionales o métodos de clasificación de referencia (baseline).

El resto del capítulo está estructurado como sigue: la sección 3.2 presenta un estado del arte que aborda los estudios principales realizados en el área y los puntos en los que la tesis hace sus principales contribuciones. La sección 3.3 presenta el procedimiento de anotación humana y discute las características del corpus y los resultados de su anotación en términos de emociones anotadas y acuerdo entre anotadores. En la sección 3.4 hay una descripción del proceso de clasificación automática de emociones, y una discusión de los resultados experimentales obtenidos para la misma. La sección 3.5 describe los enfoques estudiados con anterioridad y tiene en cuenta ambas fuentes de contexto hasta que se obtiene el enfoque óptimo. Finalmente, en la sección 3.6 hay un resumen de los beneficios de los métodos propuestos y se presentan las conclusiones extraídas de los mismos.

3.2 Estado del arte

Desafortunadamente no hay una definición consensuada de emoción. Se han realizado estudios psicológicos y biológicos a este respecto durante siglos y no es extraño encontrar referencias a Darwin o Descartes en algu-

nos artículos recientes del área. En un esfuerzo para clarificar conceptos y subrayar la dirección de las principales líneas de investigación en el área, (Cowie, 2000) distingue dos formas de interpretar las emociones: la primera es en la forma de estados discretos (p.ej. miedo, felicidad, enfado) que suelen denominarse emociones "full-blown" en la literatura; y la segunda es como un atributo de ciertos estados, que el autor denomina "estados emocionales". En el proyecto HUMAINE (Humaine emotion-research.net, 2007) también hacen esta distinción empleando otra nomenclatura: emociones episódicas y permanentes respectivamente. Los principales esfuerzos de investigación se dirigen hacia el estudio de emociones permanentes o "estados emocionales"; mientras que el principal objetivo de estudio de las emociones "full-blown" es encontrar un conjunto restringido de categorías o emociones. Existe una teoría muy extendida de que las emociones "full-blown" sólo pueden tomar unas cuantas formas fácilmente distinguibles entre sí. Sin embargo, haces esta simplificación no siempre es aplicable, puesto que para ciertas aplicaciones puede ser interesante estudiar emociones mezcladas, simuladas y/o conflictivas.

Con independencia de la definición de emoción, se pueden encontrar diversas formas de representarlas en la literatura. Pueden representarse empleando un conjunto discreto o como puntos en un espacio. En el caso continuo, las emociones se representan mediante coordenadas en el espacio con un pequeño número de dimensiones. El enfoque típico es el espacio de activación-evaluación bidimensional (Cowie et al., 2001). En el eje horizontal, la evaluación trata con la "valencia" de las emociones, esto es, evaluaciones positivas o negativas de gente, cosas o eventos. En el vertical, la activación mida la disposición del usuario a tomar una acción en lugar de no hacer nada. Las emociones "full-blown" forman un patrón circular en el espacio de activación-evaluación que ha hecho a otros autores proponer una representación en términos de ángulos y distancia al centro. Algunas herramientas como por ejemplo FEELTRACE (Cowie et al., 2000) han sido implementadas para dar una representación visual del progreso dinámico de las emociones en el círculo. Adicionalmente, los modelos 3D pueden emplearse para distinguir emociones próximas en el círculo (p.ej. miedo y enfado) mediante la inclusión de una tercera dimensión que suele ser el control percibido sobre la emoción o la inclinación a involucrarse. Las emociones se pueden representar también estructuralmente tratándolas desde un punto de vista cognitivo que describe la forma en que los usuarios tratan con la situación que causó la emoción.

Además, siguiendo la teoría de Ortony y G. L. Clore (1988), las emociones también pueden clasificarse de acuerdo con las situaciones que las provocan, según si son relacionadas con eventos, acciones de agentes o aspectos de objetos. Esta es la denominada teoría OCC que se emplea normalmente para síntesis de emociones (Zong et al., 2000; de Melo y Paiva, 2005).

Los sistemas emocionalmente inteligentes en la actualidad suelen poner énfasis bien en la causa o bien en el efecto de las emociones; pocos se centran en la gestión de las mismas. En el primer caso, la atención se centra en las razones de la aparición de la emoción, que pueden ser externas o internas al usuario; el segundo describe los efectos de dichas características en el oyente (Cowie, 2000). La investigación en estas áreas está generalmente orientada al reconocimiento en el caso de las causas y a la síntesis en los efectos. La investigación descrita en este capítulo se centra en el caso del reconocimiento.

El reconocimiento de emociones puede llevarse a cabo con métodos invasivos y no invasivos. Los métodos invasivos se basan en medidas fisiológicas cuyo cálculo suele conllevar conectar dispositivos al cuerpo del usuario. Éste es el caso del ritmo respiratorio o la conductividad de la piel (Picard, 1997). Uno de los métodos más extendidos consiste en medir la respuesta galvánica de la piel (galvanic skin response, GSR) puesto que hay una estrecha relación entre el despertar de una emoción y los cambios en la GSR (Lee et al., 2005). Otros métodos son EMG, que mide los movimientos de los músculos faciales (Mahlke, 2006), el ritmo cardíaco o más recientemente el empleo de imágenes de actividad cerebral (Critchley et al., 2005). Por el contrario, los métodos no invasivos se basan por lo general en audio y vídeo. Por una parte, el reconocimiento oral de emociones se puede llevar a cabo con información acústica y/o lingüística. La voz se ve profundamente afectada por las emociones: la acústica, el contorno, el tono, la calidad de la voz y la articulación cambian dependiendo del estado emocional, se puede encontrar un estudio muy detallado de la forma en que se presentan estos cambios en (Cowie et al., 2001). La información del lenguaje trata con las variaciones lingüísticas que se producen según el estado emocional del usuario, para lo cual ha ganado notoriedad la técnica de la representatividad emocional de las palabras (word emotional salience). Esta medida representa la frecuencia de aparición de una palabra en una categoría emocional dada, calculándose a partir de un corpus de interacciones usuario-sistema (Lee et al., 2005). Por otra parte, el recono-

cimiento por medio de vídeo normalmente presta atención a las expresiones faciales, la postura del cuerpo y los movimientos de las manos, se puede encontrar un resumen de todas ellas en (Picard y Daily, 2005). Otros autores enfatizan que las emociones se ven influenciadas por trasfondos culturales y sociales y defienden el empleo del enfoque "interaccional" (Boehner et al., 2007) junto con las medidas fisiológicas y audiovisuales.

El reconocimiento de emociones se ha empleado en el campo de la interacción hombre-máquina para diversos propósitos: en algunos dominios de aplicación es necesario reconocer el estado afectivo de los usuarios para adaptar los sistemas al mismo o incluso cambiarlo. Por ejemplo, en servicios de emergencia (Bickmore y Giorgino, 2004) o tutores inteligentes (Ai et al., 2006), es necesario saber el estado emocional del usuario para calmarlo o animarlo en actividades de aprendizaje respectivamente. Sin embargo, también hay aplicaciones para las que la gestión de emociones no es un aspecto central, pero contribuye a un mejor funcionamiento del sistema. En estos dominios, la gestión de emociones puede emplearse para resolver etapas del dialogo que causan estados emocionales negativos, así como para evitarlas y fomentar los positivos en futuras interacciones. Burkhardt et al. (2005) emplean un detector de enfado para evitar la frustración de los usuarios durante la interacción con su portal de voz. Además, las emociones no sólo son interesantes por sí mismas, sino también porque afectan al mensaje explícito que se intercambia durante la interacción: cambian las voces, las expresiones faciales, los gestos, la velocidad del discurso, etc. Este fenómeno se denomina coloreado emocional ("emotional colouring") y puede ser de gran importancia para la interpretación de la entrada del usuario. Así, en (Wahlster, 2006) emplean coloreado emocional en el contexto del sistema SmartKom para detectar sarcasmo y abordar falsas actitudes positivas.

Como se ha descrito anteriormente, el reconocimiento de emociones es un aspecto clave para obtener interacciones parecidas a las humanas. Por eso ha recibido mucha atención por parte de la comunidad científica, habiéndose contemplado desde aplicaciones en las que los cambios en el estado emocional del usuario son únicamente indicadores de que el sistema no cumple con sus expectativas, a sistemas complicados en los que las emociones son una piedra angular, como sistemas de ayuda psicológica. Esto se refleja en la cantidad de projectos internacionales e interdisciplinarles que tratan el tema, algunos de los más recientes, ordenados por fecha de comienzo, son:

- SAFIRA - Supporting Affective Interactions for Real-time Applications - Fomentando las interacciones afectivas para aplicaciones en tiempo real (The Safira Project - DFKI Page, 2002). 24 meses desde 02-05-2000 (Completado). Su propósito era el enriquecimiento de aplicaciones con una dimensión afectiva para apoyar la conducta y control afectivo de los sistemas de tiempo real multi-agente en su interacción con usuarios.

- MEGA - Multisensory Expressive Gesture Applications - Aplicaciones con gesticulación expresiva multisensorial (MEGA Project, 2001). 36 meses desde 01-11-2000 (Completado). Su propósito era el modelado, reconocimiento y síntesis de gestos en tiempo real, la comunicación en red y la interacción con contenido emocional no-verbal (p.ej. música y danza) por medio de interfaces multi-sensoriales, desde una perspectiva multimodal.

- MAGICSTER - Embodied Believable Agents - Agentes personificados creíbles (MagiCster Project Pages, 2007). 39 meses desde 01-12-2000 (Completado). Su propósito fue el diseño y evaluación de un agente conversacional personificado creíble, que hacía uso de miradas, expresiones faciales, gestos y posturas del cuerpo así como habla de forma sincronizada.

- NECA - A Net Environment for embodied emotional Conversational Agents - Un entorno de red para agentes conversacionales personificados emocionales (NECA project, 2005). 30 meses desde 01-10-2001 (Completado). Su propósito era la creación de espacios virtuales multiusuario poblados por agentes conversacionales afectivos capaces de expresarse a través de voz emocional sincronizada con expresiones no verbales.

- ERMIS - Emotionally Rich Man-Machine Interaction Systems - Sistemas de interacción hombre-máquina con riqueza emocional (EUROPA - CORDIS: Community Research and Development Information Service, 2006). 36 meses desde 01-01-2002 (Completado). Su propósito era el desarrollo de un prototipo para la interacción entre usuario y ordenador capaz de interpretar la actitud o estado emocional del usuario en términos de su discurso y/o expresiones faciales.

- PF-STAR - Preparing future multisensorial interaction research - Preparando la investigación futura en interacción multisensorial (PF-STAR home page, 2004). 24 meses desde 01-10-2002 (Completado). Su propósito era contribuir al campo de la comunicación multisensorial y multilingüe proveyendo de referencias tecnológicas, evaluaciones comparativas, y valoraciones de las posibilidades de las tecnologías clave, sobretodo las de traducción voz-a-voz, detección y expresión de estados emocionales y las tecnologías dirigidas a los niños.

- VICTEC - Virtual ICT with Empathic Characters - ICT virtual con personajes empáticos (VICTEC in Lynne Hall web page, 2005). 35 meses desde 01-03-2002 (Completado). Su propósito era el desarrollo de un toolkit que ayudase a la creación de personajes sintéticos creíbles en un entorno virtual con capacidad de establecer relaciones empáticas con niños.

- NICE - Natural Interactive Communication for Edutainment - Comunicación natural e interactiva para educación+entretenimiento (NICE project - Main page, 2007). 36 meses desde 01-03-2002 (Completado). Su propósito era alimentar al acceso interactivo natural y universal, en particular para niños y adolescentes, mediante el desarrollo de un paradigma de comunicación natural, divertida y rica, con personajes históricos y literarios.

- CHIL - Computers in the Human Interaction Loop - Ordenadores en el bucle de la interacción humana (CHIL - Computers In the Human Interaction Loop, 2007). 36 meses desde 18-12-2003 (Completado). Su propósito era la creación de entornos en los que los ordenadores sirvan a personas durante su comunicación con otras personas, momento en el que por tanto no pueden estar pendientes de la máquina.

- HUMAINE - Research on Emotions and Human-Machine Interaction - Investigación en emociones e interacción hombre-máquina (Humaine emotion-research.net, 2007). Desde 18-12-2003 (En ejecución). Su propósito es establecer las bases para el desarrollo europeo de sistemas que puedan registrar, modelar e influenciar el estado emocional de los usuarios coordinando esfuerzos para alcanzar un entendimiento común de los factores implicados.

- AMI - Augmented Multi-party Interaction (Interacción múltiple aumentada) (Augmented Multiparty Interaction Project, 2007). Desde 18-12-2003 (En ejecución). Su propósito es la creación de nuevas tecnologías multimodales que ayuden a la interacción entre humanos en el contexto de salas de reuniones inteligentes con asistentes virtuales remotos.

- INTREPID - A Virtual Reality Intelligent Multi-sensor Wearable System for Phobias' Treatment - Un sistema inteligente de realidad virtual, multi-sensor portable, para el tratamiento de fobias (Intrepid Project , A virtual reality intelligent multi-sensor wearable system for phobias' treatment). 24 meses desde 01-01-2004 (Completado).

- AUBADE - A wearable EMG Augmentation system for robust behavioural understanding - Un sistema EMG que se puede llevar puesto para una comprensión robusta del comportamiento (AUBADE, 2005). 34 meses desde 01-01-2004 (Completado). Su propósito era el desarrollo de un plataforma que se puede llevar puesta para monitorizar y reconocer ubicuamente el estado emocional de los usuarios en tiempo real, empleando señales medidas de su rostro.

- COSY - Cognitive Systems for Cognitive Assistants - Sistemas cognitivos para asistentes cognitivos (CoSy Home, 2007). 48 meses desde 01-09-2006 (En ejecución). Su propósito es la construcción de sistemas que puedan percibir, entender e interactuar con su entorno, y evolucionar para alcanzar un rendimiento similar al de una persona en actividades que requieran conocimiento específico del contexto (de la situación y la tarea).

- CALLAS - Conveying Effectiveness in Leading-Edge Living Adaptive Systems - Transmitir efectividad en sistemas adaptativos punteros (Callas - Conveying Affectiveness in Leading-Edge Living Adaptive Systems, 2007). 42 meses desde 01-11-2006 (En ejecución). Su propósito es la definición y desarrollo de una arquetiectura multimodal que incluya aspectos emocionales con el fin de dar soporte a experimentos y reutilizar y enfocar nuevas aplicaciones multimedia, sobretodo en el paradigma de la inteligencia ambiental.

- COMPANIONS - Intelligent, Persistent, Personalised Multimodal Interfaces to the Internet - Interfaces multimodales inteligentes, persistentes y personalizadas para Internet (Companions, 2007). 48 meses desde 01-11-2006 (En ejecución). Su propósito es la creación de compañeros (interfaces personalizadas y conversacionales a Internet) que conozcan a su dueño, y que funcionen en un amplio rango de plataformas, estáticas y portables, basadas en investigación integrada de alta calidad en interfaces persona-ordenador multimodales, agentes inteligentes y tecnología del lenguaje natural.

En el área del reconocimiento de emociones la mayor parte de los estudios [1] están centrados en el estudio de la adecuación de distintos clasificadores de aprendizaje máquina (Shafran y Mohri, 2005), como los k vecinos más próximos (Lee y Narayanan, 2005), modelos ocultos de Markov (Ververidis y Kotropoulos, 2006; Pitterman y Pitterman, 2006), Support Vector Machines (Morrison et al., 2007), redes neuronales (Morrison et al., 2007) o algoritmos Boosting (Liscombe et al., 2005; Forbes-Riley y Litman, 2004a). Además, se han dirigido muchos esfuerzos investigadores a encontrar las mejores características para ser clasificadas. Estas características se pueden categorizar a distintos niveles. El nivel más bajo trata con características fisiológicas que suelen medirse con métodos intrusivos como la respuesta galvánica de la piel. Los niveles acústicos y lingüísticos están más extendidos y usan características muy generalizadas que se pueden encontrar frecuentemente en la literatura como los cambios articulatorios (Cowie et al., 2001), medidas estadísticas de parámetros acústicos (Ververidis y Kotropoulos, 2006) o representatividad emocional de las palabras (Lee y Narayanan, 2005). Recientemente también se han adoptado características visuales, especialmente en sistemas multimodales, tales como las expresiones faciales, la posición del cuerpo o los movimientos de las manos (Picard y Daily, 2005; Zeng et al., 2006). Como se ha comentado anteriormente, algunos autores abogan además por introducir información acerca del fondo social y cultural de los usuarios para detectar sus emociones (Boehner et al., 2007).

Sin embargo, se le presta menos atención al proceso de entrenamiento de los algoritmos en que se basa la clasificación automática de emociones y para la cual es necesario contar con corpus anotados manualmente. Un buen

[1] Consultar Ververidis y Kotropoulos (2006) para mayor información.

esquema de anotación es esencial, pues afecta al resto de las etapas del proceso de aprendizaje. Además, la anotación manual de corpus es muy difícil, requiere mucho tiempo y es costosa, por tanto debe ser diseñada cuidadosamente. Los autores que estudian los corpus emocionales están interesados especialmente en cómo se obtienen, sobretodo en la comparación de emociones actuadas y reales (Morrison et al., 2007), sin embargo ha recibido menos atención la forma en que se debe anotar cualquiera de dichos corpus. Entre otros, Devillers et al. (2005) han propuesto líneas de trabajo para diseñar y desarrollar esquemas de anotación exitosos en términos de etiquetas y procesos de validación. Gut y Bayerl (2004) también han trabajado en medidas de fiabilidad de las anotaciones realizadas por humanos, mientras que Craggs y Wood (2003) han propuesto diversas capas de anotación de emociones.

La presente tesis va un paso más allá y estudia la manera de añadir información contextual al proceso de anotación, sugiriendo la inclusión de dos tipos de información contextual: la forma de hablar neutra de los usuarios y la historia del diálogo. La primera aporta información referente a la forma de hablar de los usuarios cuando no están expresando ninguna emoción, lo que puede llevar a un mejor entendimiento de los estados no neutros del usuario (sección 3.4.2). Mientras que la segunda implica el empleo de información acerca del estado actual del diálogo en términos de la longitud del diálogo y el número de confirmaciones y repeticiones (sección 3.4.3), lo que puede indicar de forma bastante fiable el posible estado emocional del usuario. Por ejemplo, es probable que el usuario esté enfadado si tiene que repetir la misma información en numerosos turnos consecutivos del diálogo.

En la literatura hay tres enfoques principales para recoger corpus emocionales: grabar habla espontánea, grabar emociones inducidas o usar actores que simulen las emociones. Como se muestra en la figura 3.1, en estos enfoques existe un compromiso entre la naturalidad de las emociones y el control sobre los datos recolectados: cuanto mayor sea el control sobre los datos generados, menor será la espontaneidad y naturalidad de la emoción expresada, y viceversa. Por tanto, el habla emocional espontánea, que refleja la producción completamente natural de la emoción dentro del dominio de aplicación en que se emplea el reconocedor de emociones, es el enfoque más realista. Sin embargo, se necesita un esfuerzo considerable para anotar el corpus, puesto que requiere que para cada grabación se interprete qué emoción ha sido expresada. A veces, el corpus está grabado de interacciones entere humanos

en el dominio de aplicación (Forbes-Riley y Litman, 2004a), en este caso el resultado también es natural pero no es directamente aplicable al caso en que los usuarios humanos interactúan con una máquina. En el otro extremo, el habla emotiva actuada es más facil de manipular y evita la necesidad de anotación, pues las emociones mostradas en cada grabación se conocen de antemano. Los resultados obtenidos dependen en gran medida de las habilidad de los actores, con lo que sólo se obtienen buenos resultados con actores con suficiente preparación dramática. Cuando se emplean actores no expertos, es necesaria otra fase para descartar las grabaciones que no reproducen apropiadamente la emoción requerida. En un punto medio se encuentran las emociones inducidas, que pueden ser más naturales como aquellas inducidas durante el transcurso de un videojuego (Johnstone, 1996), o más fáciles de manipular, como las que son inducidas al hacer leer a los sujetos textos estrechamente relacionados con emociones específicas (Stibbard, 2000).

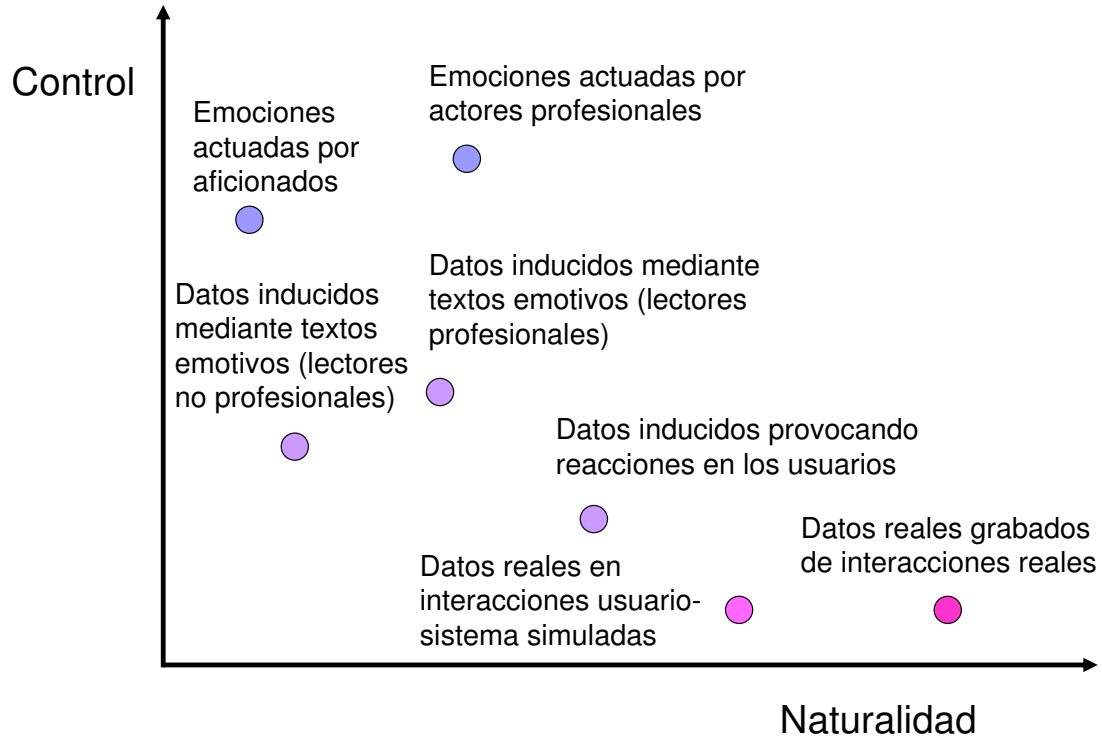


Figura 3.1. *Naturalidad vs. control en los enfoques principales para la generación de corpus multimodales*

Como algunos autores han indicado, p.ej. Douglas-Cowie et al. (2003), la relación entre las emociones actuadas y las espontáneas no se conoce de forma exacta. Sin embargo, como indica Johnstone (1996), incluso el habla actuada profesionalmente pierde realismo, pues existen algunos efectos que no pueden ser controlados de forma consciente. Diversos estudios han mostrado que no es apropiado emplear datos actuados para reconocer emociones que se dan de forma natural (Vogt y André, 2005; Wilting et al., 2006).

Como el objetivo de investigación futuro es construir un reconocedor de emociones para el sistema de diálogo UAH (capítulo 6), y tendría que trabajar con emociones que tienen lugar en tiempo real, para los experimentos se empleó un corpus de emociones reales extraídas de interacciones de los usuarios con el sistema. Las emociones reales no inducidas son difíciles de encontrar en los corpus en español. Por ejemplo, de los 70 corpus estudiados por Douglas-Cowie et al. (2003) y Ververidis y Kotropoulos (2006), sólo hay tres en español: González (1999), Montero et al. (1999) y Iriondo et al. (2000). Tal y como se muestra en la tabla 3.1, dos de ellos se emplearon para la síntesis de emociones en lugar de para su reconocimiento. La tabla muestra asimismo un conjunto de corpus en español empleados para el estudio de características emocionales (Adell et al., 2005) y estudios de propósito general (Hozjan et al., 2002). Ninguno de ellos se adquirió a partir de interacciones reales, siendo ocho el número máximo de autores empleados, mientras que el corpus UAH se adquirió a partir de las interacciones de más de 60 usuarios reales.

3.3 Anotación humana del corpus UAH

La anotación de emociones es una tarea altamente subjetiva dado que una misma elocución puede ser percibida por anotadores diferentes como distintas emociones. La manera más fiable de obtener anotaciones rigurosas es reclutar a anotadores especializados, como por ejemplo psicólogos que estén entrenados en el reconocimiento de emociones humanas. Desafortunadamente, en la mayoría de los casos los anotadores expertos son difíciles de encontrar y por tanto la anotación debe hacerse con anotadores inexpertos. En este caso, todos los anotadores eran no expertos pues no habían recibido ninguna preparación específica para el reconocimiento de emociones.

Referencia	Sujetos	Finalidad	Tipo
González (1999)	-	Reconocimiento	Inducida
Montero et al. (1999)	1 actor	Síntesis	Simulada
Iriondo et al. (2000)	8 actores	Síntesis	Simulada
Hozjan et al. (2002)	2 actores	Estudio, síntesis y reconocimiento	Simulada
Adell et al. (2005)	1 actriz, 1 lector profesional, 1 miembro del Parlamento Español	Estudio de las características de las emociones	2 simulados, 1 natural
UAH corpus	Aprox. 60 usuarios del sistema UAH	Reconocimiento	Natural

Tabla 3.1. *Corpus emocionales en español*

Para conseguir la mejor anotación posible empleando anotadores no expertos, el proceso de etiquetado debe diseñarse rigurosamente. Vidrascu y Devillers (2005) sugieren diversas etapas para decidir la lista de etiquetas y el esquema de anotación, reglas de segmentación, número de anotadores, procesos de validación y estudios de consistencia.

La primera etapa consiste en decidir las etiquetas a emplear para la anotación. Nuestro principal interés radica en el estudio de estados emocionales negativos de los usuarios, principalmente para detectar frustración debido al mal funcionamiento de los sistemas. Por tanto, la clasificación se ha realizado entre las tres emociones negativas con mayor presencia en el corpus UAH: *enfado*, *aburrimiento* y *duda*. Para la anotación del corpus se empleó además una cuarta categoría: *neutro*, que representa un estado emocional no negativo (las emociones positivas se han tratado como neutras). La categoría neutra ha sido empleada sólo para la anotación humana del corpus, el resto de experimentos se centran exclusivamente en la distinción entre las emociones negativas consideradas.

Se tomó la decisión de utilizar un número grande e impar de anotadores: nueve, que es más de lo que normalmente se encuentra en los estudios previos, p.ej. Forbes-Riley y Litman (2004a) y Lee y Narayanan (2005). En cuanto al "tamaño del segmento" (*segment length*) al que asignar emocio-

nes, se ha optado por emplear las elocuciones completas, ya que no era útil usar unidades menores como podrían haber sido las palabras, puesto que el objetivo perseguido en la tesis es analizar la emoción como una respuesta completa a las intervenciones del sistema, sin considerar los posibles cambios emocionales dentro de las elocuciones.

En el proceso de anotación descrito, el corpus fue anotado dos veces por cada anotador: en primer lugar de forma ordenada y en segundo de forma desordenada. De la primera forma, los anotadores contaban con información acerca del contexto del diálogo y la manera de hablar de los usuarios. En el segundo caso, los anotadores no tenían dicha información, así que sus anotaciones se basaban meramente en información acústica acerca de la elocución actual.

La emoción final asignada a cada elocución en los esquemas ordenados y desordenados era la indicada por la mayoría de los anotadores, en las situaciones en las que no había una emoción mayoritaria (p.ej. 4 *neutro*, 4 *aburrimiento* y 1 *duda*), se le daba prioridad a las no neutras (en el ejemplo anterior *aburrimiento*). Las anotaciones finales para todo el corpus se realizaron teniendo en cuenta las 18 anotaciones, lo que permitía resolver posibles conflictos entre dos emociones no neutras (p.ej. 4 *duda*, 4 *aburrimiento* y 1 *neutro*).

3.3.1. Cálculo del nivel de acuerdo entre anotadores

Se han empleado diversos coeficientes Kappa para el estudio del grado de acuerdo entre anotadores para ambos estilos de anotación (ordenado y desordenado). Los coeficientes Kappa se basaban en la idea de calcular el cociente de la proporción de parejas de anotadores en acuerdo (P_o) con la proporción esperada de anotadores que están de acuerdo por casualidad (P_c). De esta forma, se obtiene el nivel de acuerdo no fortuito observado ($P_o - P_c$) respecto a todos los acuerdos no fortuitos posibles ($1 - P_c$):

$$\kappa = \frac{P_o - P_c}{1 - P_c} \tag{3.1}$$

En la tesis se han empleado cuatro coeficientes Kappa distintos (figura 3.2) con lo cuales se estudiaron principalmente dos cuestiones: i) el impacto de las tendencias de anotadores (annotator bias), es decir, dado un número

fijo de acuerdos, el efecto que tiene la distribución de desacuerdos entre categorías sobre el valor de los coeficientes Kappa; y ii) el nivel de importancia de todos los desacuerdos posibles en la tarea, puesto que el desacuerdo entre emociones que son fácilmente distinguibles debería tener un impacto más negativo en el valor del coeficiente Kappa que los desacuerdos entre categorías muy diferentes.

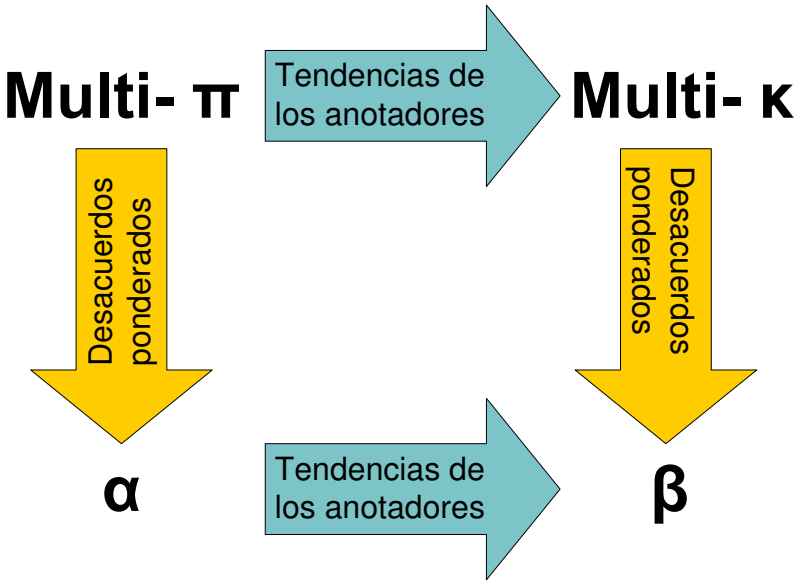


Figura 3.2. *Coeficientes Kappa empleados en la experimentación*

El coeficiente Kappa más sencillo empleado es el propuesto por Fleiss (1971) como una generalización para múltiples anotadores del coeficiente π de Scott para dos anotadores (Scott, 1955). Ha habido mucha confusión en la literatura acerca del Kappa de Fleiss, pues muchos autores lo han presentado en su lugar como una generalización del κ de Cohen (Cohen, 1960). Este tema se discute con más detalle en (Artstein y Poesio, 2005), quienes han hecho un esfuerzo considerable para clarificar las definiciones de los distintos coeficientes Kappa. Para evitar inconsistencias, se ha seguido su notación para todos los coeficientes Kappa empleados en el capítulo. En particular, el Kappa de Fleiss se nota como multi-π.

El cálculo de multi-π se basa en la ecuación 3.1, donde el acuerdo observado (P_o) se calcula como el número de casos en que dos anotadores distintos se ponen de acuerdo en anotar una elocución particular con la misma

emoción:

$$P_o = \frac{1}{UA(A-1)} \sum_{u=1}^{U} \sum_{e=1}^{E} n_{ue}(n_{ue} - 1) \tag{3.2}$$

En la ecuación 3.2, U es el número de elocuciones (utterances) a ser anotadas, A el número de anotadores, E el número de emociones, y n_{ue} el número de veces que la elocución 'u' ha sido anotada con la emoción 'e'.

Fleiss asume que todos los anotadores comparten la misma distribución de probabilidad. Esto implica que la probabilidad de que un anotador clasifique una elocución 'u' con una emoción 'e' en particular, puede calcularse como la probabilidad global de anotar 'u' como 'e'. Esta probabilidad global se ha calculado como el número total de asignaciones a la emoción 'e' realizadas por todos los anotadores (n_e en la ecuación 3.3) dividida por el número de total de asignaciones ($U \cdot A$). El acuerdo fortuito (ecuación 3.3) se ha calculado como la probabilidad de que cualquier par de anotadores anoten la misma elocución con la misma emoción, que se asumió como la probabilidad conjunta de que cada uno de ellos hiciese esta asignación de forma independiente, pues los anotadores juzgaron todas las elocuciones de forma independiente los unos de los otros.

$$P_c^{\pi} = \sum_{e=1}^{E} \left(\frac{1}{UA} n_e \right)^2 \tag{3.3}$$

El cálculo de multi-π asume que cada anotador sigue la misma distribución global de elocuciones en emociones. Sin embargo, dicha simplificación puede no ser plausible en cualquier dominio debido al efecto de las denominadas "tendencias de los anotadores" (*annotator bias*) en el valor de Kappa. En los experimentos realizados, dichas tendencias pueden definirse como la medida en que los anotadores difieren en la proporción de emociones dado un número particular de acuerdos. Con el resto de los parámetros fijados, el valor de Kappa crece al hacerlo dichas tendencias, esto es, cuando las proporciones de desacuerdos no son iguales para todas las emociones y hay un mayor desequilibrio entre ellas. Esta es la denominada *segunda paradoja del Kappa*. Podemos encontrar diferentes estudios de su impacto en la literatura, p.ej. Feinstein y Cicchetti (1990), Cicchetti y Feinstein (1990), Lantz y Nebenzahl (1996), y Artstein y Poesio (2005).

Para estudiar si la inclusión de distintos comportamientos de anotación mejora los valores de Kappa, se ha calculado la Kappa de Davies y Fleiss (1982), que se ha notado como multi-κ, siguiendo la nomenclatura de Artstein y Poesio (2005). Como pasa con multi-π, el cálculo de multi-κ también se basa en la ecuación 3.1 y tiene el mismo acuerdo observado (ecuación 3.2). Sin embargo, para el acuerdo fortuito, incluye una distribución distinta para cada anotador. Por tanto, en este caso la probabilidad de que un anotador 'a' clasifique una elocución 'u' con la emoción 'e' se calcula con el número observado de elocuciones asignadas a la emoción 'e' por ese anotador (n_{ae}), dividido por el número total de elocuciones (U). La probabilidad de que dos anotadores se pongan de acuerdo en anotar una elocución 'u' con la emoción 'e' es de nuevo la probabilidad conjunta de que cada uno haga dicha anotación de forma independiente:

$$P_c^{\kappa} = \frac{1}{\binom{A}{2}} \sum_{e=1}^{E} \sum_{j=1}^{A-1} \sum_{k=j+1}^{A} \frac{n_{a_j e}}{U} \frac{n_{a_k e}}{U} \tag{3.4}$$

A pesar de incluir diferencias entre los anotadores, multi-κ da a todos los desacuerdos la misma importancia. En la práctica, todos los desacuerdos no son igualmente probables y no tienen el mismo impacto sobre la calidad de los resultados de notación. Por ejemplo, en nuestros experimentos, el desacuerdo entre *neutro* y *enfado* es más fuerte que entre *neutro* y *duda*, puesto que el primero se da entre dos categorías más fácilmente distinguibles.

Para tener en cuenta toda esta información, se han empleado coeficientes Kappa ponderados (Cohen, 1968; Fleiss y Cohen, 1973), que se centran en los desacuerdos en lugar de en los acuerdos[2]. Su cálculo se basa en la ecuación 3.5 (equivalente a la ecuación 3.1):

$$\kappa_w = 1 - \frac{\overline{P}_o}{\overline{P}_c} \tag{3.5}$$

donde $\overline{P}_o$ representa el desacuerdo observado y $\overline{P}_c$ el desacuerdo fortuito. Para todos los coeficientes empleados, el desacuerdo observado ha sido calculado como el número de veces que la elocución 'u' ha sido anotada con dos emociones distintas e_j and e_k por cada par de anotadores, ponderado según

[2]Se puede encontrar un cálculo alternativo basado en acuerdo en (Sim y Wright, 2005)

la distancia entre las emociones:

$$\overline{P}_o = \frac{1}{UA(A-1)} \sum_{u=1}^{U} \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} n_{ue_j} n_{ue_k} distancia(e_j, e_k) \qquad (3.6)$$

Por consiguiente, el cálculo de los coeficientes ponderados implica emplear medidas de distancia entre las cuatro emociones empleadas para la anotación (*neutro*, *enfado*, *aburrimiento* y *duda*). Para hacerlo, la lista discreta de emociones ha sido dispuesta en un espacio continuo, concretamente el de activación-evaluación (Russell, 1980). Como se ha comentado anteriormente, las emociones forman un patrón circular en dicho espacio, por lo que algunos autores han propuesto una representación de las mismas en forma de ángulos y distancia al centro. Aprovechando esta disposición circular, para el cálculo de los coeficientes Kappa ponderados se han empleado las distancias angulares entre las emociones estudiadas. En lugar de establecer nuestro propio emplazamiento de las emociones en el círculo, hemos empleado las disposiciones empleadas por el estudio seminal (Plutchik, 1980), con el fin de evitar errores de medida. La lista de Plutchik cuenta con 40 emociones y sus ángulos respectivos y ha sido ampliamente aceptada por la comunidad científica. En esta lista se contemplaban explícitamente aburrido (136,0º) y enfadado (212,0º), pero no dubitativo. Las emociones más parecidas eran "vacilante" (uncertain), "desconcertado" (bewildered) y "confuso" (confused), que sólo se diferenciaban en 2º dentro del círculo. De entre ellos se escogió "inseguro"(139,3º) por ser la emoción que mejor reflejaba el estado emocional que buscábamos anotar. Sin embargo, otros autores como Scherer (2005) han considerado explícitamente dubitativo como un estado emocional. Plutchik (1980) no reflejó el neutro en su lista, puesto que no es una emoción sino la "ausencia" de ella, en su lugar empleó un estado denominado "aceptando" (accepting) como punto de partida del círculo (0º), que se ha empleado como *neutro* en nuestros experimentos.

La distancia entre las cuatro categorías se calculó en grados teniendo en cuenta el ángulo que cada una formaba en el círculo. Se consideró el ángulo menor en todos los casos (x or 360-x) de forma que la distancia entre cada dos ángulos estuviese siempre entre 0 y 180 grados. Para el cálculo de los coeficientes Kappa, las distancias se convirtieron en pesos con valores

entre 0 y 1. Un peso 0 (que corresponde con una distancia de 0º en el enfoque propuesto) implica anotar la misma emoción y por tanto no caer en desacuerdo. Por el contrario, peso=1 (distancia de 180º) se corresponde con anotaciones completamente contradictorias y por tanto máximo desacuerdo. Las distancias y pesos resultantes aparecen en la tabla 3.2.

Ángulo / Peso	Neutro	Enfadado	Aburrido	Dubitativo
Neutro	0,00º/0,00	148,00º/0,82	136,00º/0,75	139,30º/0,77
Enfadado	148,00º/0,82	0,00º/0,00	76,00º/0,42	72,70º/0,40
Aburrido	136,00º/0,75	76,00º/0,42	0,00º/0,00	3,30º/0,02
Dubitativo	139,30º/0,77	72,70º/0,40	3,30º/0,02	0º/0,00

Tabla 3.2. *Distancia entre las emociones*

No existe consenso en la comunidad científica acerca de las propiedades de las medidas de distancia. Sin embargo, Artstein y Poesio (2005) han propuesto algunas restricciones: la distancia de una categoría consigo misma debe ser mínima y la distancia entre dos categorías no debe depender del orden (es decir, que la distancia de A a B debe ser igual que la de B a A). Como puede observarse por la simetría de la tabla, las medidas de distancia empleadas en la tesis siguen dichas restricciones:

- El ángulo que una emoción forma consigo misma es 0º

$$\forall e \in E, distancia(e, e) = 0$$

- El ángulo entre una emoción A y otra B es el mismo en ambas direcciones (pues se ha establecido escoger el ángulo mínimo):

$$\forall e_A, e_B \in E, distancia(e_A, e_B) = distancia(e_B, e_A)$$

Como puede observarse en la tabla, las mayores distancias se dan entre no neutros y neutros. Por tanto, al calcular los coeficientes Kappa, se adjudicó mayor importancia a aquellos desacuerdos en los que un anotador juzgó una elocución como neutra y otro anotador la estimó como no neutra, considerándose menos relevantes los desacuerdos entre emociones no neutras.

Se han calculado dos coeficientes Kappa ponderados: α de Krippendorff (2003) y β de Artstein y Poesio (2005). Ambas comparten el mismo cálculo para el desacuerdo observado (ecuación 3.5), para α el desacuerdo fortuito se calcula:

$$\overline{P}_c^{\alpha} = \frac{1}{UA(UA-1)} \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} n_{e_j} n_{e_k} distancia(e_j, e_k) \qquad (3.7)$$

Como puede observarse en la ecuación 3.7, este coeficiente no considera las tendencias de los anotadores. Esto se resuelve con el coeficiente β en el que se mide además el comportamiento observado de cada anotador:

$$\overline{P}_c^{\beta} = \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} \left[\frac{1}{U^2 \binom{A}{2}} \sum_{m=1}^{A-1} \sum_{n=m+1}^{A} n_{a_m e_j} n_{a_n e_k} distancia(e_j, e_k) \right] \qquad (3.8)$$

Los resultados para cada coeficiente descrito se listan en la tabla 3.3 y se discuten en la siguiente sección.

Coeficiente	Desordenado	Ordenado
multi-π	0,3256	0,3241
multi-κ	0,3355	0,3256
α	0,3382	0,3220
β	0,3393	0,3237

Tabla 3.3. *Valores de los coeficientes Kappa para los distintos esquemas de anotación*

3.3.2. Discusión de los resultados de la anotación humana

Como se ha comentado previamente, una de las dificultades del reconocimiento de emociones en sistemas de diálogo oral es que en la mayoría de los dominios de aplicación los corpus obtenidos están muy desequilibrados,

debido a que normalmente una proporción mayor de neutros que de elocuciones emocionales (Forbes-Riley y Litman, 2004a; Morrison et al., 2007). Esto está en concordancia con los resultados experimentales obtenidos en la tesis puesto que se anotaron más del 85,00 % de las elocuciones como *neutro* en media entre los nueve anotadores. También se ha observado que esta proporción se ve afectada en un 3,40 % de los casos por el estilo de anotación. En concreto, con el estilo ordenado, el 87,28 % fueron anotadas como *neutro*, mientras que para la anotación desordenada, el corpus estaba más desequilibrado aún; el 90,68 % de las elocuciones fueron anotadas como *neutro*. La figura 3.3 muestra la proporción de emociones no neutras anotadas por los nueve anotadores. Como puede observarse, el estilo de anotación ordenado obtuvo un mayor porcentaje de aburridos; el 39,00 % más que en el caso desordenado. La figura tambiéen muestra que la categoría *enfado* se ve afectada substancialmente por el estilo de anotación, sin embargo, la categoría *duda* parece ser prácticamente independiente del estilo de anotación.

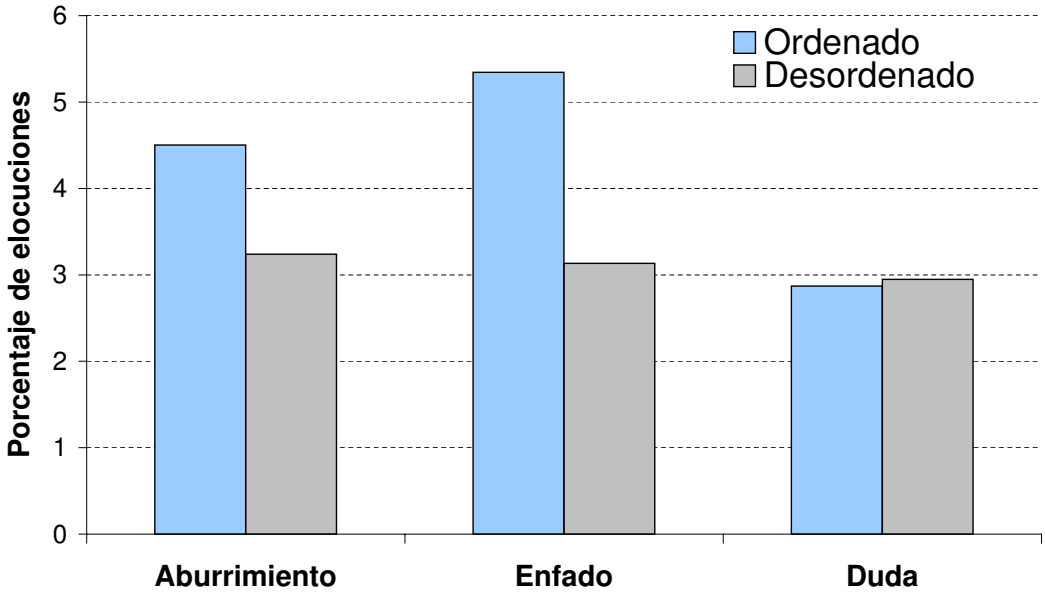


Figura 3.3. *Proporción de elocuciones anotadas como no neutras*

Es posible que estos resultados se deban a que la incorporación de contexto en el caso ordenado empuje a los anotadores a clasificar elocuciones correspondientes a los mismos diálogos en las mismas categorías emocionales. De esta forma, no hay transiciones notables entre elocuciones consecutivas. Por ejemplo, si se detecta enfado en una elocución es probable que la si-

guiente también sea anotada como *enfado*. Además, el contexto permite a los anotadores tener información acerca de la forma de hablar del usuario y la historia del diálogo. Por el contrario, en el caso desordenado los anotadores sólo tenían información sobre la elocución actual. Por tanto, a veces no podían discriminar si el usuario estaba enfadado o si normalmente hablaba fuerte y rápido. De este modo, se trata de un hecho relevante a tener en cuenta cuando la anotación la realizan anotadores no expertos, que el método más común, barato y rápido. Además, al escuchar el corpus en el caso ordenado, los anotadores tenían información sobre de la posición del turno de usuario actual dentro del diálogo, lo que daba una pista de su estado emocional. Por ejemplo, es más probable que el usuario se aburra en un diálogo largo, o que se enfade después de muchas peticiones de confirmación generadas por el sistema.

Como puede observarse en la tabla 3.3, los valores de los distintos coeficientes Kappa varían ligeramente dependiendo del esquema de anotación empleado. En el caso desordenado tanto tener en cuenta las tendencias de anotadores (multi-κ vs. multi-π y β vs. α) como ponderar los desacuerdos (β y α vs. multi-κ) mejoran los valores de acuerdo. Sin embargo, en el caso ordenado solamente mejora los valores de acuerdo tener en cuenta las tendencias de anotadores mientras que los desacuerdos ponderados reducen el valor de Kappa. Esto se deriva del incremento de anotaciones no neutras discutido anteriormente. Teniendo en cuenta que la gran mayoría de acuerdos se producen cuando los anotadores anotan la misma elocución como neutra (como se observa en la figura 3.4), un incremento en el número de emociones anotadas como no neutras provoca más discrepancias entre anotadores y, por tanto, reduce el valor de Kappa.

Igualmente, como puede observarse en la figura 3.5 la mayoría de los desacuerdos se producen entre categorías neutras y no neutras, que son las emociones con mayor distancia en nuestro esquema de ponderación (tabla 3.2), provocando que los acuerdos ponderados sean menores en el caso del esquema ordenado.

Para estudiar el efecto de las tendencias en anotadores, se realizó un cálculo exhaustivo de los acuerdos entre cada par de anotadores. Como se observa en la figura 3.5 no hubo anotadores que tuvieran un acuerdo significativamente pobre con el resto. Sin embargo, cuando se examinaron los resultados de anotación, se encontró que había diferencias notables entre

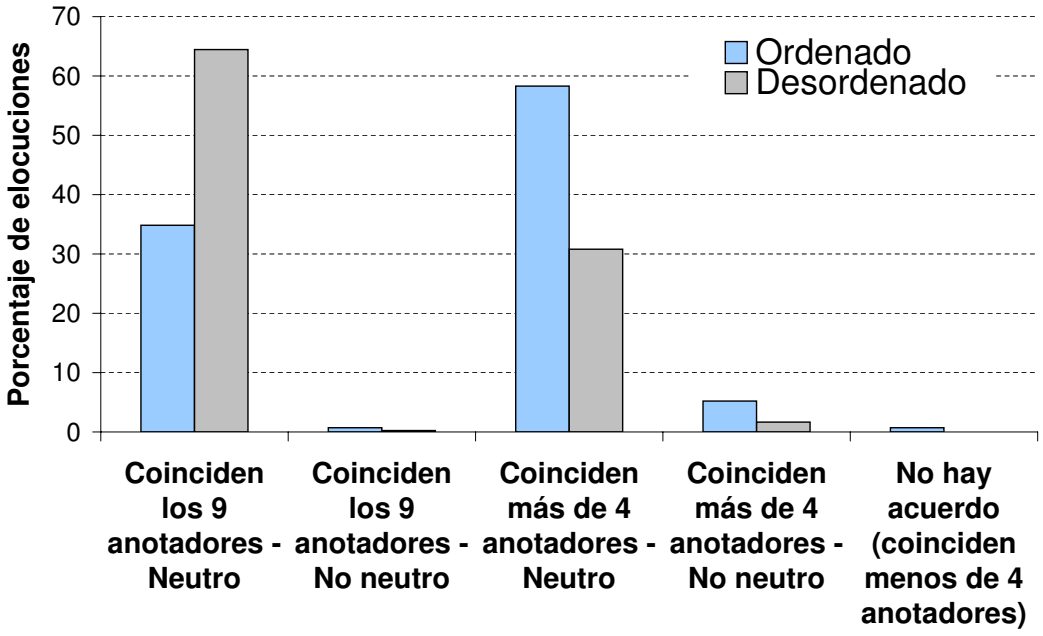


Figura 3.4. *Porcentaje de acuerdo entre los anotadores*

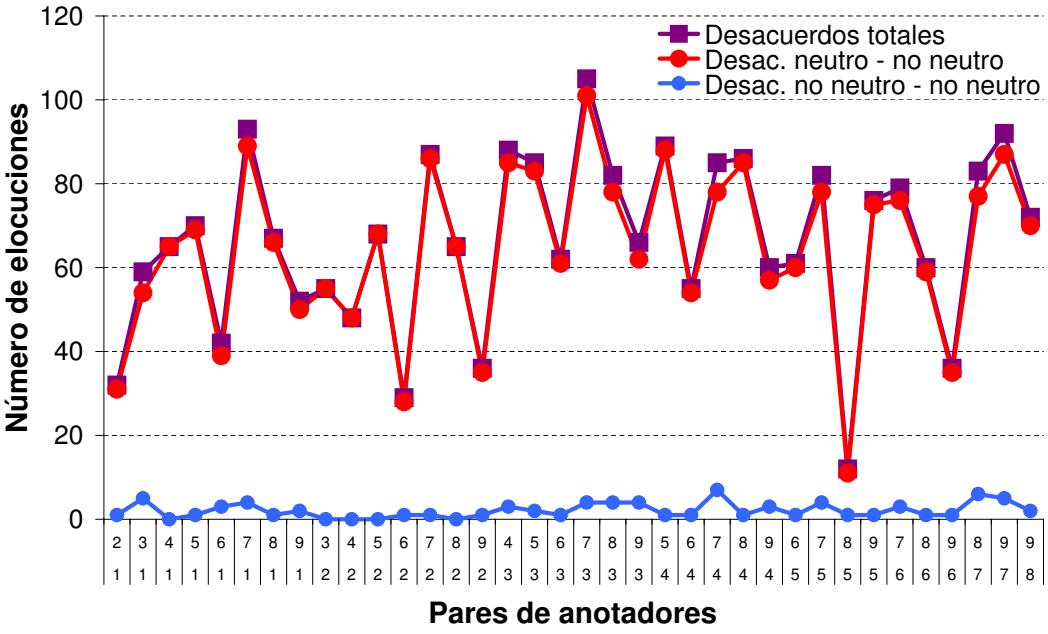


Figura 3.5. *Desacuerdos entre cada par de anotadores al seguir el esquema ordenado*

aquellos anotadores que estaban habituados al acento andaluz (en el que estaban pronunciadas las elocuciones) y aquellos que no lo estaban. Como se ha explicado previamente, el corpus fue grabado a partir de interacciones con el sistema UAH, cuyos usuarios eran principalmente estudiantes y profesores de la Universidad de Granada, por lo que su forma de hablar estaba influenciada por el acento andaluz que se caracteriza entre otras cosas por un ritmo más rápido y una mayor fuerza expiratoria (Gerfen, 2002; O'Neill, 2005). Del grupo de anotadores empleados, seis estaban habituados al acento andaluz (anotadores 1, 2, 3, 4, 6 y 9 en la figura 3.5) y tres no (anotadores 5, 7 y 8).

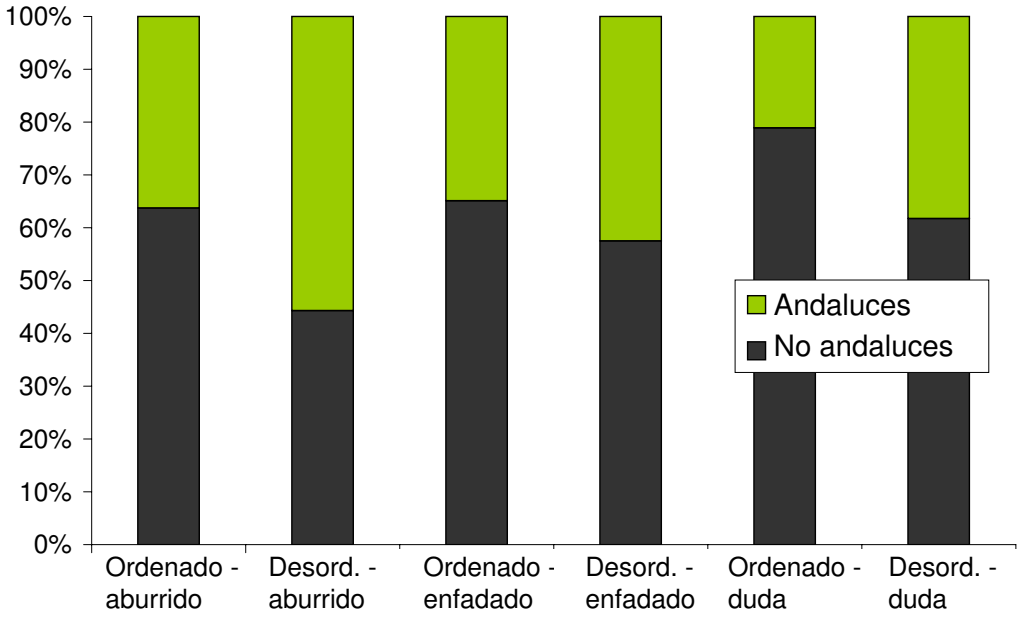


Figura 3.6. *Proporción de emociones anotadas por cada uno de los tipos de anotadores*

La figura 3.6 muestra, para el número de total de anotaciones realizadas en cada categoría, qué porcentaje corresponde a cada tipo de anotador. Como puede observarse, en todos los casos menos uno (especialmente en aquellos obtenidos mediante el esquema ordenado), los anotadores no habituados al acento andaluz marcaron alrededor de un 60 % de las emociones encontradas para esa característica emocional. Esto se debe a la confusión de características del acento con atributos emocionales, como por ejemplo, el ritmo rápido con un indicativo de enfado. El efecto de los esquemas de anotación en ambos tipos de anotadores ha sido estudiado y los resultados

se muestran en la tabla 3.4.

	Anotadores andaluces		Anotadores no andaluces	
	Desordenado	Ordenado	Desordenado	Ordenado
multi-π	0,3608	0,3234	0,3734	0,5593
multi-κ	0,3621	0,3275	0,3746	0,5598
α	0,3595	0,3248	0,3644	0,5691
β	0,3607	0,3265	0,3703	0,5697

Tabla 3.4. *Valores Kappa para los distintos tipos de anotadores*

Como se puede observar en la tabla, los anotadores acostumbrados al acento andaluz obtuvieron valores de Kappa muy similares para ambos esquemas de anotación (variando de 0,3234 a 0,3621). Para estos anotadores, los valores de Kappa eran menores porque anotaron menos elocuciones como neutras. Por el contrario, los anotadores acostumbrados al acento andaluz obtuvieron valores de Kappa diferentes dependiendo del esquema de anotación empleado: en el caso ordenado los valores variaban de 0,5593 a 0,5697 mientras que en el desordenado entre 0,3639 y 0,3746. Esto se debe a un mayor decremento del acuerdo fortuito. Como se muestra en la figura 3.7, el acuerdo observado era más o menos constante, mientras que el fortuito disminuyó drásticamente en el esquema ordenado.

La razón más probable para este decremento es el número de neutros anotados por los anotadores no acostumbrados al andaluz. Esto ocurre en ambos esquemas de anotación, pero el número de neutros anotados es mayor para el caso desordenado y es por esto que los resultados son más parecidos a los obtenidos para los anotadores andaluces con el esquema de anotación ordenado. Aunque el número de anotaciones no neutras aumentó proporcionalmente al decremento en las neutras, el desequilibrio del corpus conllevó a que la probabilidad de estar de acuerdo por casualidad en la categoría neutra fuese más importante en el cálculo del acuerdo fortuito global. Por ejemplo, en el caso de multi-κ, el acuerdo fortuito (P_c) se calculó como la sumatoria del acuerdo fortuito en cada emoción: ($P_c = P_c^{neutro} + P_c^{aburrimiento} + P_c^{enfado} + P_c^{duda}$). Los valores de acuerdo fortuito cuando los anotadores que no estaban acostumbrados al acento andaluz empleaban el esquema ordenado fueron:

- $P_c^{neutro} = 0{,}6645$,
- $P_c^{aburrimiento} = 0{,}0052$,
- $P_c^{enfado} = 0{,}0069$ y
- $P_c^{duda} = 0{,}0008$.

Para el resto de los anotadores fueron:

- $P_c^{neutro} = 0{,}8137$,
- $P_c^{aburrimiento} = 0{,}0010$,
- $P_c^{enfado} = 0{,}0014$ y
- $P_c^{duda} = 0{,}0008$.

Por tanto, P_c^{neutro} fue el factor determinante en la obtención del P_c global.

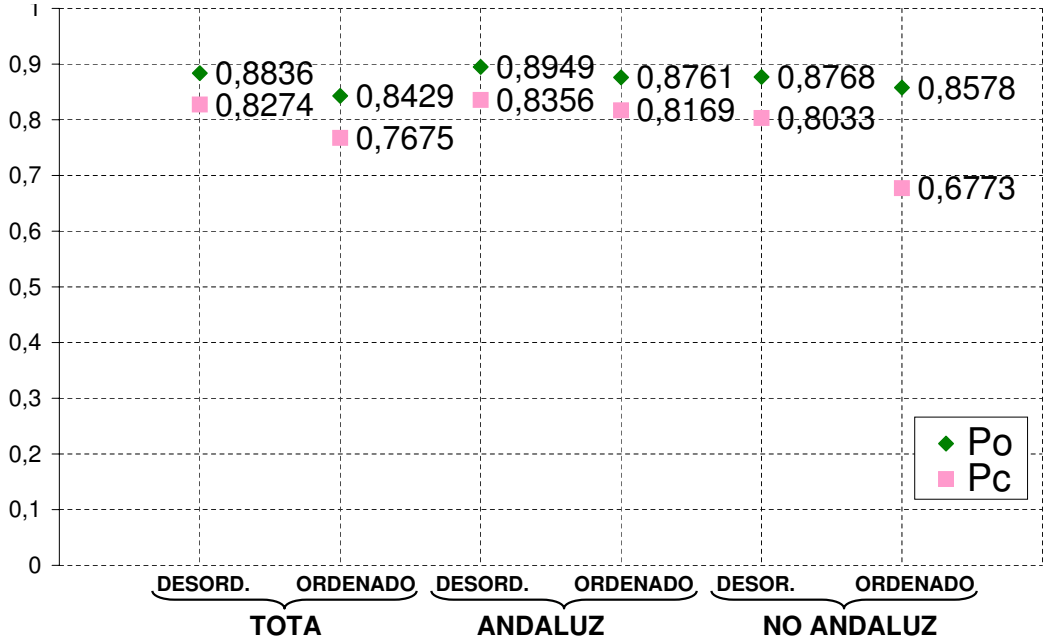


Figura 3.7. *Valores de acuerdo observado y fortuito para multi-κ*

La situación en la que a pesar de tener un número idéntico de acuerdos, la distribución de éstos entre las categorías de anotación afecta profundamente a los coeficientes Kappa se conoce como la *primera paradoja del Kappa*. Este fenómeno establece que, siendo el resto de parámetros iguales, el valor de Kappa crece con distribuciones simétricas de los acuerdos. Es decir, que si la prevalencia de una categoría sobre las demás es muy alta, entonces el acuerdo fortuito (P_c) también es alto y la Kappa decrece considerablemente (Feinstein y Cicchetti, 1990; Cicchetti y Feinstein, 1990).

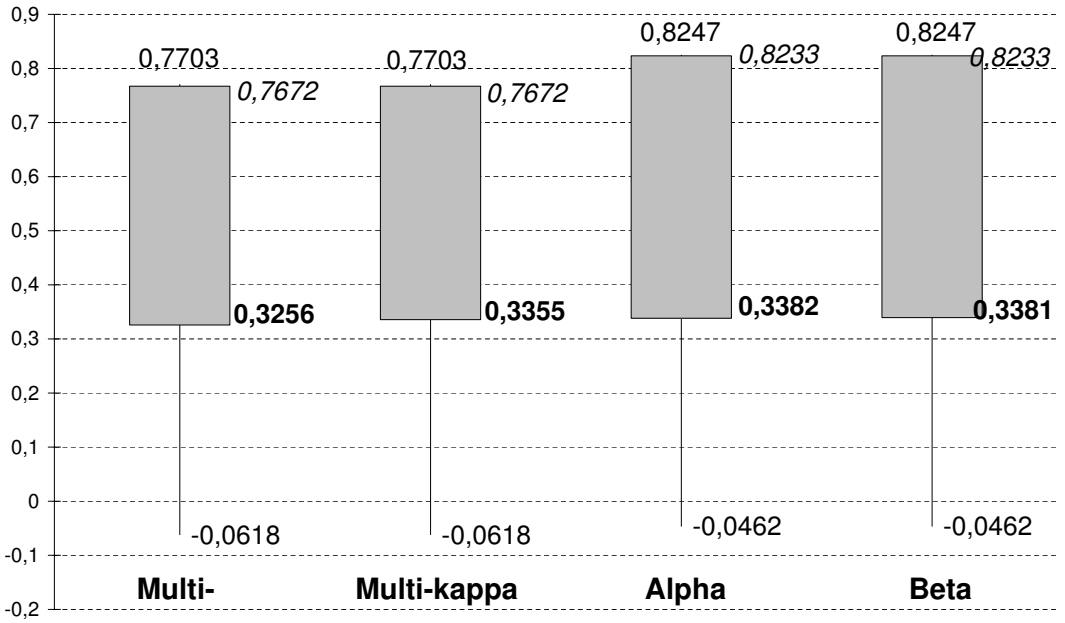
Por tanto, la primera paradoja del Kappa puede afectar enormemente a los valores Kappa y por consiguiente debe ser tenida en cuenta en su interpretación. No existe una interpretación universalmente aceptada para los valores de los coeficientes Kappa. Una de las más globalmente empleadas es la presentada por Landis y Koch (1977), que realiza una correspondencia entre intervalos de valores Kappa e interpretaciones de los acuerdos. Siguiendo este enfoque, los resultados experimentales de la tesis indican un nivel de acuerdo aceptable ("fair agreement") para ambos esquemas de anotación y con los cuatro coeficientes Kappa estudiados. Alternativamente Krippendorff (2003) estableció 0,65 como umbral para la aceptabilidad de los resultados de acuerdo, en este caso, nuestro valor mayor para Kappa (0,3393) no sería aceptable. Sin embargo, la mayoría de los autores parecen coincidir en que emplear unos intervalos pre-establecidos de valores Kappa no ofrece la suficiente información como para realizar una interpretación justificada de la aceptabilidad de los resultados de acuerdo obtenidos. Con el fin de conseguir un marco de trabajo más completo, algunos autores como Dunn (1989), proponen poner el Kappa en contexto aportando sus valores *máximo*, *mínimo* y *normal*, que pueden ser calculados a partir del acuerdo observado (P_o) como sigue (Lantz y Nebenzahl, 1996):

$$kappa_{max} = \frac{P_o^2}{(1 - P_o)^2 + 1} \tag{3.9}$$

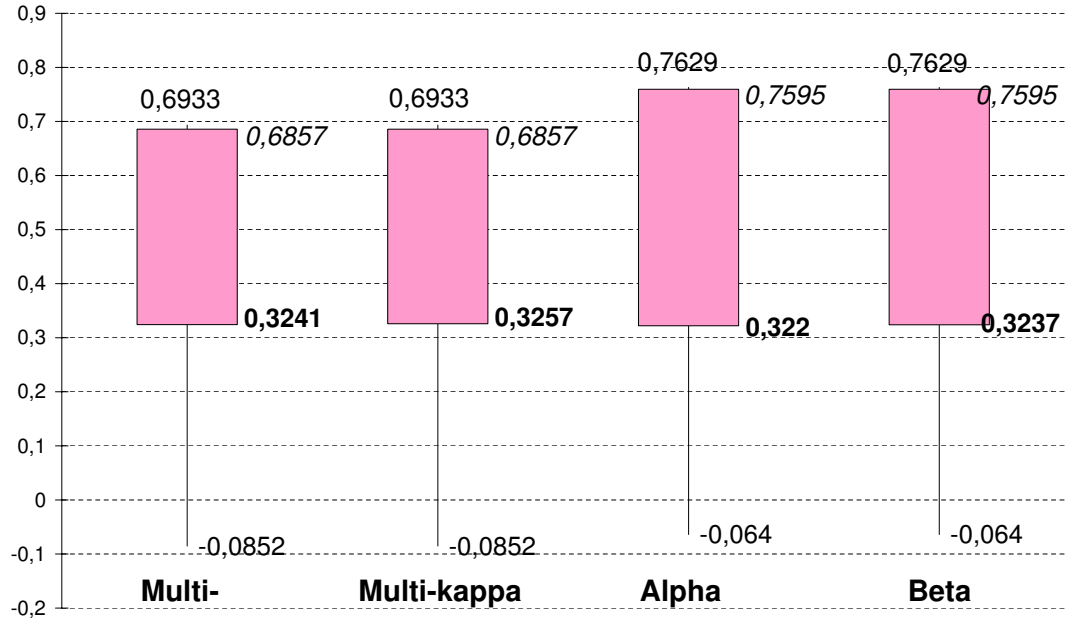
$$kappa_{min} = \frac{P_o - 1}{P_o + 1} \tag{3.10}$$

$$kappa_{nor} = 2P_o - 1 \tag{3.11}$$

Los resultados Kappa obtenidos se presentan junto con sus valores $kappa_{max}$, $kappa_{min}$ y $kappa_{nor}$ en la figura 3.8.



(a) Esquema desordenado



(b) Esquema ordenado

Figura 3.8. *Valores Kappa máximo, mínimo, normal (cursiva) y observados (negrita)*

Tal y como se observa en la figura, para el mismo acuerdo observado, los valores posibles de Kappa pueden variar mucho de $kappa_{min}$ a $kappa_{max}$ dependiendo del equilibrio del corpus. El valor $Kappa_{max}$ se obtiene cuando se desequilibran al máximo los desacuerdos mientras se mantienen equilibrados los acuerdos, mientras que $kappa_{min}$ se obtiene cuando los acuerdos están desequilibrados y los desacuerdos equilibrados. $Kappa_{nor}$ no corresponde a un valor ideal de Kappa, sino con distribuciones simétricas tanto de acuerdos como de desacuerdos. Puede observarse en la figura cómo el desplazamiento entre los valores obtenidos y los normales es más pequeño con el esquema ordenado (figura 3.8(b)). Por tanto, este esquema no sólo permite el reconocimiento de más emociones no neutras, sino también la obtención de valores de Kappa que, aunque sean más pequeños que en el caso desordenado en valor absoluto, están más cerca de los valores de acuerdo normales y máximos alcanzables y más lejos del mínimo.

Como describen Lantz y Nebenzahl (1996), los desplazamientos respecto al valor $kappa_{nor}$ indican asimetría en los acuerdos o desacuerdos dependiendo de si están más cerca del valor mínimo o máximo respectivamente. En la figura 3.8, el desplazamiento entre los valores Kappa observados y los normales se representa con un rectángulo. Los resultados corroboran que presentar los valores Kappa es más informativo cuando se ponen en contexto, pues se obtiene un indicativo valioso de los posibles desequilibrios que, a nuestro parecer, deben ser considerados para llegar a conclusiones apropiadas sobre la fiabilidad de las anotaciones. Así, en los experimentos descritos se dieron desplazamientos significativos respecto a $kappa_{nor}$ en todos los casos (rectángulos), lo que corrobora que había una gran asimetría entre categorías. Esto se debe al fenómeno de "prevalencia" anteriormente comentado (primera paradoja del Kappa).

Como se ha comentado, la prevalencia aparece como una consecuencia inevitable del desequilibrio natural de los corpus emocionales no actuados, donde la categoría neutra es claramente predominante. Por consiguiente, queda demostrado que los enfoques basados únicamente en valores preestablecidos de aceptabilidad (Landis y Koch, 1977; Krippendorff, 2003) no son apropiados para este dominio de aplicación. Algunos autores ya han aportado medidas adicionales para complementar la información presentada por los coeficientes Kappa. De esta manera, Forbes-Riley y Litman (2004a) aportan tanto los valores Kappa como los del acuerdo observado, mientras que

Lee y Narayanan (2005) muestran los valores Kappa junto con un test de hipótesis.

Aunque los valores Kappa obtenidos en reconocimiento de emociones empleando corpus desequilibrados suelen ser bajos, por ejemplo de 0,32 a 0,42 en Shafran et al. (2003) y por debajo 0,48 en Ang et al. (2002) y Lee y Narayanan (2005), no se encuentra en la literatura ningún estudio detallado acerca de la problemática de los coeficientes Kappa en este área, ni tan siquiera en artículos dedicados explícitamente a los retos de la anotación de emociones (Devillers et al., 2005). Además, incluso cuando se aportan otros valores de acuerdo junto con los coeficientes Kappa (Forbes-Riley y Litman, 2004a; Lee y Narayanan, 2005), sólo se calcula un coeficiente Kappa (usualmente multi-π) y no se argumenta por qué existe tanta diferencia entre los valores Kappa y las medidas aportadas. Por tanto, el estudio presentado es uno de los primeros en tratar los coeficientes Kappa en el área de la anotación de emociones no actuadas así como los fenómenos relacionados con su interpretación.

Finalmente, para obtener una idea más aproximada sobre el nivel de acuerdo real alcanzado entre los nueve anotadores, exponemos a continuación los valores del acuerdo observado en la tabla 3.5, que ha sido empleada junto con los valores de Kappa por otros autores en diferentes áreas de estudio (Ang et al., 2002; Forbes-Riley y Litman, 2004a). Como se observa en la tabla, en todos los casos el acuerdo observado se sitúa por encima de 0,85. Esta medida no contempla el efecto de la prevalencia (ver figura 3.7), y por tanto los valores no son mayores en el caso de los anotadores no habituados al acento andaluz cuando seguían el esquema ordenado.

De los resultados anteriores, es posible concluir que emplear el esquema ordenado ha permitido la anotación de más emociones no neutras. Desafortunadamente esto se traduce en valores menores de los coeficientes Kappa pues la mayoría de los acuerdos se producían para los neutros. Esto indica que se deben emplear múltiples anotadores para anotar las emociones naturales y obtener corpus emocionales fiables. Una alternativa para abordar el problema de la alta probabilidad de acuerdos fortuitos, consiste en maximizar el acuerdo observado. Por ejemplo, Litman y Forbes-Riley (2006) proponen emplear lo que denominan "anotación consensuada" (*consensus labelling*), es decir, alcanzar consenso entre los anotadores hasta obtener un 100 % de acuerdo observado.

Esquema de anotación	Tipo de anotador	Acuerdo observado	Acuerdo observado ponderado
Desordenado	Total	0,8836	0,9117
	Andaluz	0,8950	0,9197
	No andaluz	0,8767	0,9050
Ordenado	Total	0,8429	0,8800
	Andaluz	0,8761	0,9049
	No andaluz	0,8578	0,895

Tabla 3.5. *Acuerdo observado para todos los esquemas de anotación y tipos de anotadores*

En nuestro caso, los valores Kappa calculados han sido útiles para comparar los dos esquemas de anotación. Como se muestra en la figura 3.8, aunque los valores de Kappa y del acuerdo observado son menores en el caso ordenado, se ha encontrado que este tipo de anotación es de gran utilidad para obtener valores más cercanos a los máximos alcanzables. También se desprende del estudio presentado en la tesis que evaluar la fiabilidad de los procesos de anotación donde los acuerdos están tan desequilibrados puede conducir a valores muy bajos de los coeficientes Kappa (tabla 3.3) que están lejos de los altos valores de acuerdos observados (tabla 3.5). Ha sido necesario incluir otras fuentes de información como el acuerdo observado y los valores de Kappa mínimos, máximos y normales, junto con los coeficientes Kappa, con el fin de obtener interpretaciones significativas. Además, como se muestra en la tabla 3.5, dar un peso a los distintos tipos de desacuerdo puede incrementar considerablemente el acuerdo observado entre los anotadores y se ha presentado un método para calcular las distancias entre estos desacuerdos.

3.4 Clasificación automática del corpus UAH

Los resultados experimentales descritos en la sección 3.3 muestran que la información contextual es muy importante para los anotadores humanos. En esta sección se estudia si la discriminación automática entre emociones también se ve afectada por este factor. El interés específico en esta tesis es la distinción de emociones negativas y por tanto sólo se han empleado las ca-

tegorías no neutras como entrada a los algoritmos de aprendizaje. La razón para esto no es reducir el efecto del desequilibrio del corpus, sino llevar a cabo un estudio más profundo de las diferencias entre las emociones negativas consideradas. La línea de futuro más inmediata consistirá en añadir un reconocedor de neutros vs. no-neutros con el fin de construir un reconocedor que maneje este desequilibrio natural (ver capítulo 6). Los experimentos presentados en esta sección pueden clasificarse en dos tipos: relacionados con la voz y con el diálogo. En el primer grupo, el aprendizaje máquina se ha empleado para la distinción de las tres emociones de interés (*enfado*, *aburrimiento* y *duda*) y se han medido los beneficios de emplear un enfoque novedoso propuesto para la normalización acústica con el fin de obtener mejores resultados de clasificación. Para el segundo grupo se ha considerado adicionalmente el conocimiento acerca del contexto de la interacción, presentando asimismo un método de clasificación y de representación del contexto.

Con fines comparativos, el primer enfoque usado es un "baseline" que siempre anota cada elocución de usuario con la misma emoción independientemente de la entrada. En el corpus UAH la categoría emocional más frecuente es *enfado*, por lo que el baseline siempre anota cada elocución con esta etiqueta. El segundo algoritmo empleado es un perceptrón multicapa (Multilayer Perceptron, MLP) (Rumelhart et al., 1986; Bishop, 2006), para el que se ha empelado una topología con una capa oculta con número de nodos igual a la suma del número de características y el número de emociones dividida por dos. La tasa de aprendizaje, que determina la velocidad de convergencia de la búsqueda, se estableció a 0,3 para evitar que fuese demasiado grande (en cuyo caso podría perder mínimos u oscilar abruptamente) o demasiado pequeña (provocando una convergencia lenta). Para prevenir la sobre-clasificación, las pasadas sobre los datos se restringieron a 500. Adicionalmente, se estableció un umbral de validación de 0,2 para determinar las veces consecutivas que el error del conjunto de validación se deterioraba antes de parar el entrenamiento. Para mejorar el rendimiento, se introdujo un momentum de 0,2 para el aprendizaje de los pesos que configuraban el MLP. Para entrenar el MLP y llevar a cabo la experimentación se empleó el WEKA toolkit (Witten y Frank, 2005). Es una colección de algoritmos de aprendizaje automático de código abierto para el pre-procesamiento, clasificación, regresión, agrupamiento y visualización de datos.

Para entrenar y evaluar los algoritmos de aprendizaje propuestos, se ha empleado una técnica de validación cruzada con cinco particiones (5-fold cross-validation). Los experimentos consistieron en cinco intentos donde el corpus se dividió aleatoriamente en cinco subconjuntos aproximadamente iguales (20 % del corpus cada uno). Cada intento empleaba por turnos cada una de las particiones para la evaluación (testing) y el resto (80 % del corpus) para entrenamiento (training), de forma que después de los cinco intentos cada instancia había sido empleada una vez para evaluación. Además, se extrajo una participación de "tuning" (20 %) para cada partición de entrenamiento con el fin de hacer la selección de características. La evaluación se llevó a cabo en dos fases. En la primera, los algoritmos se entrebanaban con el 60 % de las elocuciones de entrenamiento y se evaluaban con el 20 % empleado para tuning. En la segunda fase, la partición de entrenamiento completa se empleaba para entrenar el MLP, y la parte de test (20 %) para evaluación. Con fines comparativos, este segundo paso se llevó a cabo, por una parte, empleando todas las características de las elocuciones de entrenamiento, mientras que por otra parte, también se emplearon únicamente las características seleccionadas en el primer paso.

Finalmente, para todos los experimentos descritos en esta sección, se comprobó la significatividad de los resultados empleado un corrected paired t-tester disponible en la herramienta de análisis del Weka 3.5.4 Experimenter[3] (Witten y Frank, 2005), con significatividad 0,05.

3.4.1. Clasificación automática basada en características acústicas estándar

Para estos experimentos se emplearon 60 características, que incorporan aquellas empleadas normalmente en estudios previos (Devillers et al., 2005; Lee y Narayanan, 2005; Morrison et al., 2007). Éstas son estadísticas a nivel de elocución correspondientes a los cuatro grupos dispuestos en la tabla 3.6.

[3]El valor del estadístico t se calcula en Weka como:

$$t = \frac{\overline{d}}{\sqrt{\left(\frac{1}{k}+\frac{n_2}{n_1}\right)\sigma_d^2}}$$

En nuestro caso, con una validación cruzada en 5 particiones repetida 5 veces: $k = 25$, $\frac{n_2}{n_1} = \frac{0,2}{0,8}$ y σ_d^2 es la varianza en 25 diferencias.

Categoría	Características
Frecuencia fundamental (F0)	Min, max, rango, media, mediana, desviación típica, pendiente, coef. correlación, error regresión, valor en el primer segmento con voz, valor en el último segmento con voz
F1, F2, B1, B2	Min, max, rango, media, valor medio en el primer segmento con voz, valor en el último segmento con voz
Energía	Min, max, rango, media, mediana, desviación típica, pendiente, coef. correlación, error regresión, valor en el primer segmento con voz, valor en el último segmento con voz
Ritmo	Tasa, duración con voz, duración sin voz, valor en el primer segmento con voz, número de segmentos sin voz

Tabla 3.6. *Características acústicas empleadas para la clasificación*

El primer grupo está compuesto de características del tono. El tono depende de la tensión de las cuerdas vocales y de la presión del aire subglotal (Ververidis y Kotropoulos, 2006), y puede emplearse para obtener información acerca de las emociones en la voz. Como señala Hansen (1996), los valores medios del tono pueden usarse como indicadores significativos del habla emocional cuando se comparan con las condiciones neutras. Todas las características del tono en la porción de voz del discurso han sido calculadas. Concretamente, se presta atención al valor mínimo, máximo, medio, mediano, a la desviación estándar, al valor en el primer y último segmento con voz, al coeficiente de correlación, la inclinación y el error en la regresión lineal que describe la línea que se ajusta al contorno del tono. Todos los parámetros de duración han sido normalizados respecto a la duración de la elocución para obtener resultados comparables entre todas las elocuciones del corpus. Para extraer el tono se ha empleado el algoritmo modificado de autocorrelación propuesto por Boersma (1993).

El segundo grupo está compuesto de características relacionadas con los dos primeros formantes (F1 y F2) y sus anchos de banda (B1 y B2). Únicamente se emplearon los dos primeros formantes porque ha sido demostrado empíricamente que añadir información acerca de una tercera frecuencia no introduce ningún conocimiento adicional ni con corpus de emociones reales ni tampoco con actuadas (Morrison et al., 2007). Estas frecuencias son una representación de las resonancias del tracto vocal: al hablar cambia la configuración del tracto vocal para distinguir los fonemas que se quieren pronunciar, originando cambios en los formantes. Distintas formas de hablar producen variaciones en las posiciones típicas de los formantes, en particular, para el caso de emociones en el habla el tracto vocal se modifica con cada estado emocional. Como apunta Hansen (1996), hablantes estresados o deprimidos no articulan los sonidos con el mismo esfuerzo que en estados neutros. Las características que se han empleado para las categorías F1, F2, B1 y B2 son valor mínimo, valor máximo, rango, media, mediana, desviación típica y valor en el primer y último segmento con voz de cada elocución.

La energía se considera como tercer grupo de características. Como comentan Ververidis y Kotropoulos (2006), ésta puede emplearse para el reconocimiento de emociones porque está relacionada con la fase de aparición de las emociones. La variación de energía en las palabras o elocuciones es un indicador significativo para varios estilos de habla, puesto que el esfuerzo vocal y el ratio en la duración de los partes con voz y sin voz cambia. Hansen (1996) demostró que el enfado aumenta significativamente la intensidad y por tanto la energía. Para estas características, sólo se han empleado los valores no nulos de energía, al igual que pasaba con el tono, obteniendo los valores mínimo y máximo, la media, la mediana, la desviación típica, el valor en el primer y último segmento con voz, la correlación, la pendiente y el error de la regresión lineal de la energía.

El cuarto grupo está compuesto por características rítmicas. Éstas están basadas en la duración de los segmentos con voz y sin voz. Un segmento se considera sin voz (unvoiced) si su frecuencia fundamental es cero. Esto ocurre porque F0 equivale a la frecuencia fundamental de los pulsos glotales, que sólo se generan en presencia de discurso. Las características de ritmo y duración puede ser buenos indicadores de emoción, tal y como han mostrado estudios previos. Así, Boersma (1993) indicó que la varianza en la duración decrece en la mayoría de los dominios bajo condiciones de estrés. Se han

calculado cinco características relacionadas con el ritmo: tasa de discurso, duración de segmentos con voz, duración de segmentos sin voz, duración del segmento con voz más largo y número de segmentos sin voz. Todas estas características se han normalizado con la duración global de la elocución. La tasa de discurso se ha calculado como el número de sílabas normalizado entre la duración de la elocución. Para calcular la duración de la elocución se ha empleado el número de frames multiplicado por el tamaño del frame.

Empleando las 60 características acústicas descritas y la estrategia de validación 5x5, se ha obtenido una tasa de reconocimiento del 35,42 % con el MLP y del 51,67 % con el baseline que siempre clasifica la emoción mayoritaria. Los estudios de significatividad empleando un t-test con significativdad 0,05 han mostrado que esta diferencia es significativa.

No todas las características empleadas para la clasificación son necesariamente informativas. Las características innecesarias hacen el proceso de aprendizaje más lento e incrementan la dimensionalidad del problema. Por tanto, se ha realizado un proceso de selección de características (Guyon y Elisseeff, 2003) empleando tres métodos: en primer lugar, un "forward selection algorithm" como el empleado por Lee y Narayanan (2005), que seleccionó el valor B1 en el último segmento con voz y la energía máxima. En segundo lugar, un método genético de búsqueda que comenzaba con un conjunto vacío de atributos y usaba 20 generaciones y una probabilidad de 0,6 en el cruce y 0,033 en la mutación. Las características seleccionadas por este método fueron: F1 máximo, F1 mediana, B1 mínimo, B1 rango, B1 mediana, B1 en el último segmento con voz, B2 mínimo, B2 máximo, B2 mediana, máximo de energía, rango de la energía y energía en el último segmento con voz. En tercer lugar, los atributos fueron ordenados utilizando un ranking basado en la ganancia de información ("information gain"), cuyos resultados fueron: energía máxima (clasificada con 0,50), B1 en el último segmento con voz (con 0,46) y el resto de características con 0. Teniendo en cuenta las tres aproximaciones, el subconjunto óptimo estaba compuesto por B1 en el último segmento con voz y energía máxima. Al clasificar solamente con las características seleccionadas, no se obtuvieron mejoras en la experimentación, puesto que el porcentaje de elocuciones correctamente clasificadas fue 49,00 %, que es peor que la tasa de acierto obtenida para el baseline. Sin embargo esta diferencia no es significativa según el t-test, lo que indica que los resultados para el MLP y el baseline son equivalentes tras la selección.

3.4.2. Clasificación automática basada en características acústicas normalizadas

Para reproducir la información acerca de la forma de hablar de los usuarios que tenían los anotadores humanos con el esquema ordenado, se propone un enfoque en el que las características acústicas se normalizan alrededor de la voz neutra del usuario. Por ejemplo, si un usuario 'A' siempre habla muy alto y rápido, y un usuario 'B' siempre habla de forma relajada, algunas características acústicas del estado neutro de 'A' pueden coincidir con las de 'B' enfadado, lo que podría hacer fallar a la clasificación automática para uno de los usuarios. Esto es lo que ocurrió con los anotadores que no estaban acostumbrados al acento andaluz (sección 3.3.2), que confundían el ritmo rápido y el alto volumen de los usuarios con cuyas interacciones se recopiló el corpus.

Con el fin de llevar a cabo la normalización propuesta se han calculado las características neutras de cada usuario en cada diálogo y substraído de los valores obtenidos para el resto de las elocuciones de dicho usuario. Para calcular la voz neutra, se podría haber empleado el valor medio de todas las elocuciones de dicho usuario anotadas como *neutro* en el corpus por los anotadores. Sin embargo, el objetivo de la investigación en la tesis es que sea directamente aplicable para su uso real en el sistema UAH y es imposible llevar a cabo este cálculo en tiempo de ejecución pues supondría conocer todas las elocuciones de dicho usuario con antelación. Por tanto, se ha considerado la primera elocución del usuario como neutra, asumiendo que se encuentra inicialmente en un estado emocional no negativo. Además, el interés de la tesis radica sólo en las emociones causadas por la interacción con el sistema, asumiendo que el usuario se encuentra en un estado neutro cuando comienza la interacción con el mismo. Esta licencia es posible en dominios que no están directamente relacionados con situaciones altamente afectivas, este es el caso de los sistemas de información y de reserva que son las aplicaciones típicas en las que se emplean los sistemas de diálogo orales. Para estos dominios de aplicación, el número de diálogos en los que el usuario está de antemano en un estado emocional negativo es despreciable.

La precisión obtenida con el MLP empleando las características normalizas fue 53,17 %. Por consiguiente, introducir contexto acústico permitió al MLP mejorar los resultados del baseline, sin embargo la mejora no fue significativa. Empleando las características descritas en la sección 3.4.1 (B1 en

el último segmento con voz y energía máxima) se obtuvo un 69,33 % de elocuciones correctamente clasificadas, que mostró ser signficativo con el t-test. En el caso no normalizado, la selección de características no obtuvo ninguna mejora. De este modo, emplear características acústicas normalizadas llevó a una mejora del 17,66 % (tasa de reconocimiento del 69,33 %) sobre el baseline, siendo éste además el que mejor resultado obtuvo en el caso de la clasificación no normalizada (figura 3.9).

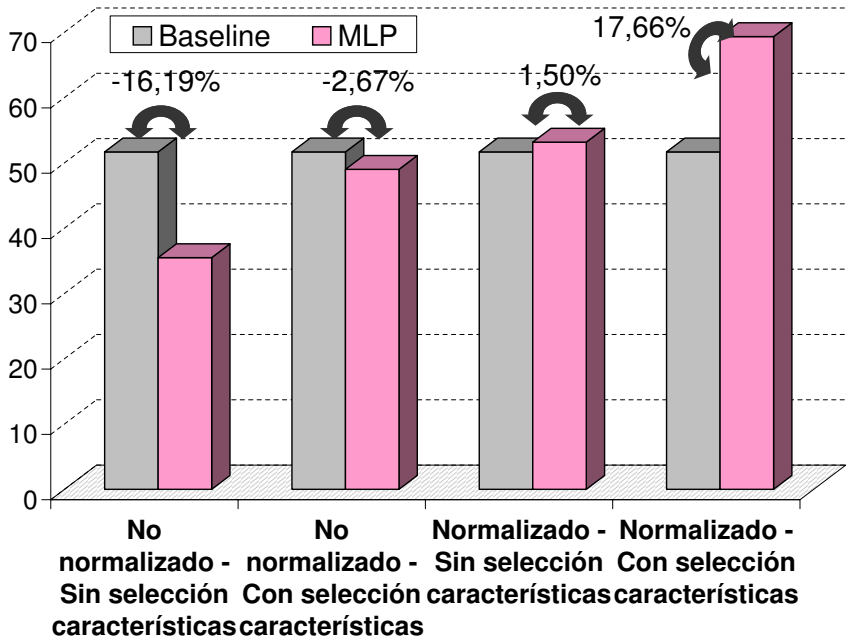


Figura 3.9. *Tasa de acierto para enfado, aburrimiento y duda según si se consideran características acústicas normalizadas o no normalizadas, así como selección o no de características*

Por tanto, la normalización de las características tradicionales lleva a un incremento considerable en el porcentaje de emociones correctamente reconocidas con respecto al baseline. Este es un resultado muy importante ya que, a pesar del desequilibrio natural de los corpus de emociones no actuadas, se pueden obtener altas tasas de reconocimiento cuando directamente se asigna la emoción más frecuente a todas las elocuciones. En nuestro caso, el baseline arrojaba una tasa del 50,00 %. Sólo con la normalización el MLP pudo obtener mejores resultados que el baseline, que se mejoraron en un 17,66 % al usar selección de características acústicas.

Un estudio de las matrices de confusión con clasificación normalizada y sin normalizar mostró que la categoría *duda* se solía confundir con *enfado* o *aburrimiento* con porcentajes de confusión por encima del 20 % en la mayoría de los casos. Con la anotación humana se obtuvo un resultado similar cuando se vio que el esquema ordenado no mejoraba la anotación de *duda* (como se observa en la figura 3.3). Estos resultados muestran que la información contextual afecta al reconocimiento automático del habla al emplear estos métodos de clasificación de forma parecida a como afecta a la anotación humana.

Así que, para distinguir *duda* del resto de las emociones negativas, son necesarias fuentes adicionales de información contextual. La propuesta de la tesis es reconocer automáticamente las tres emociones empleado un método en dos pasos. En el primer paso, se emplea información acústica y el contexto de la voz neutra del usuario para distinguir entre *enfado* y *dudaOaburrimiento*. En el segundo, el contexto del diálogo se usa para discernir entre *duda* y *aburrimiento*.

En el primer paso, el procedimiento de normalización descrito anteriormente se empleó para reconocer *enfado* vs. *dudaOaburrimiento*. Para optimizar los resultados, se llevó a cabo otra selección de características en el que las características óptimas eran aquellas que mejor discriminaban entre esas dos categorías. Empleando los mismos algoritmos de selección de características, se obtuvo un subconjunto compuesto por tres características: mediana de F2, energía máxima y energía mediana. Los resultados obtenidos se muestran en la figura 3.10. Todos demostraron ser significativamente mejores que el baseline según el t-test, exceptuando el primer caso (sin normalizar y sin selección de características), donde los resultados eran del mismo orden para el MLP y el baseline. El mejor resultado para *enfado* y *dudaOaburrimiento* se alcanzó con selección de características en el esquema ordenado, con el que se obtuvo una precisión del 80,00 %.

Con estos experimentos se ha mostrado que las características acústicas normalizadas con el contexto de la voz neutra del usuario son preferibles a las no normalizadas, pues permiten una mejora del 17,66 % (tasa de reconocimiento del 69,33 % frente al 51,67 %) cuando se reconocen las tres emociones negativas, tal y como se muestra en la figura 3.9. Además, si la información acústica se emplea como un primer paso de reconocimento que distingue entre *enfado* y *dudaOaburrimiento*, se puede alcanzar una precisión

del 80,00 %, lo que representa una mejora del 28,33 % sobre el baseline, y del 11,75 % sobre el caso en que no se emplea información contextual sobre la voz neutra del usuario (figura 3.10). En la siguiente sección, se describe el segundo paso en el que se añade el contexto del diálogo para distinguir entre *duda* y *aburrimiento*.

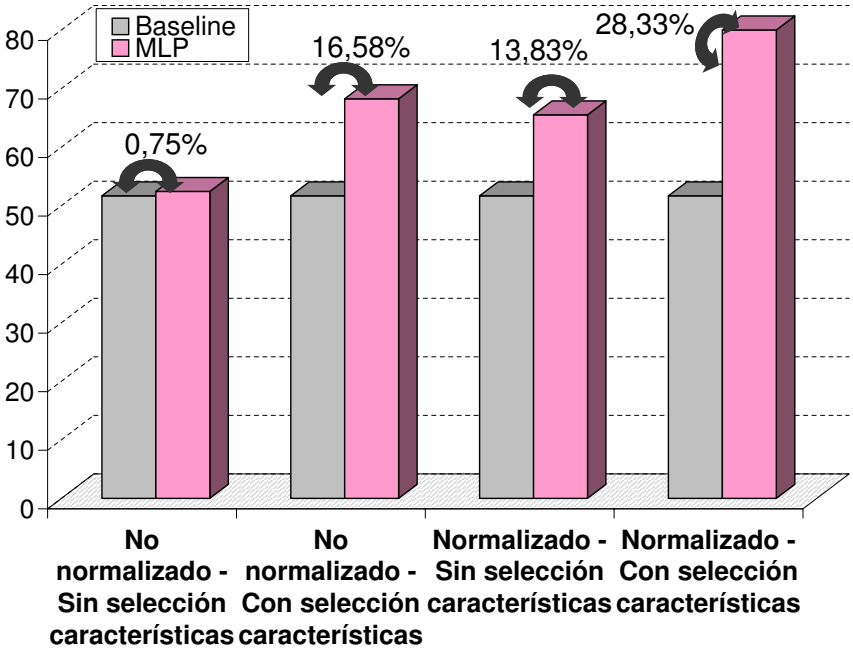


Figura 3.10. *Tasa de acierto para enfado y dudaOaburrimiento según si se consideran características acústicas normalizadas o no normalizadas, así como selección o no de características*

3.4.3. Clasificación automática basada en el contexto del diálogo

Como se describió en la sección 3.3, el contexto del diálogo se puso a disposición de los anotadores humanos en el esquema de anotación ordenado. Para representar esta información contextual en el caso del reconocimiento automático, se han empleado dos etiquetas: *profundidad* y *anchura*. La primera de ellas indica el número total de turnos del diálogo, mientras que la segunda denota el número de turnos adicionales necesarios para obtener un dato en particular (p.ej. el apellido de un profesor). Otros autores han es-

tudiado el uso de la estructura del discurso de forma similar en otras áreas. Rotaru y Litman (2006) estudiaron la forma en que localizaciones específicas dentro de la estructura del discurso son más propensas a errores de reconocimiento del habla en sistemas de diálogo. Para ello, cuantifican la posición de los turnos de usuario empleando dos medidas similares a las propuestas en la tesis: "profundidad" y "transición". Sin embargo, su enfoque se basa en modelos intencionales del diálogo, que lo consideran dividido en diferentes objetivos o intenciones que tienen que satisfacerse para completar una tarea. Por consiguiente, definen la "profundidad" de un turno de usuario como el número de subdiálogos de intención diferentes (o sub-objetivos) en la historia previa del diálogo. Por lo que en su enfoque, varios turnos de usuario pueden tener la misma "profundidad". De forma parecida, su "transición" captura la posición en el discurso relativa al turno previo, que está descrita empleando etiquetas para los distintos tipos de transiciones entre el turno actual y el anterior (p.ej. si el turno actual introduce un nuevo objetivo o continúa con el anterior).

Aunque también están basadas en posiciones "verticales" y "horizontales" de los turnos de usuario en la estructura del diálogo, el enfoque propuesto en la tesis emplea representaciones numéricas, lo que permite el cálculo automático del contexto del diálogo a partir de la historia del diálogo, sin introducir anotaciones manuales adicionales como las etiquetas de "transición" del enfoque de Rotaru y Litman. Así que, se propone una nueva forma de calcular el contexto del diálogo en la que los distintos valores de la componente vertical (*profundidad*) no están compartidos por los distintos turnos. Únicamente la etiqueta *anchura* está relacionada con el subdiálogo en que se encuentra el usuario. Sin embargo, el interés para la presente tesis doctoral está únicamente en los tipos de subdiálogo que pudieran causar una emoción negativa, que han sido considerados aquellos en los que el usuario no aporta nueva información sino que tiene que repetir, corregir o confirmar datos.

El cálculo del contexto del diálogo en el enfoque propuesto se realiza como sigue:

- La *profundidad* se inicializa a 1, dado que 0 significaría que el usuario ha colgado antes de decir nada. La *profundidad* se incrementa en 1 con cada nuevo turno del usuario y cada vez que la interacción vuelve hacia atrás (p.ej. al menú principal).

- La *anchura* se inicializa a 0, lo que indica que el usuario no necesita repetir ninguna información al sistema inicialmente. Se incrementa en 1 por cada turno del usuario generado para confirmar o repetir datos o pedir ayuda al sistema, también se incrementa en 1 si el usuario tiene que desambiguar alguna entrada, incluso cuando aporte nueva información. La *anchura* vuelve a establecerse a 0 cuando el usuario introduce un dato complemente nuevo.

Para ilustrar el método seguido, a continuación se expone un ejemplo de diálogo con el sistema UAH (S=sistema, U=usuario):

S1> Bienvenido al sistema UAH. ¿En qué puedo ayudarle?
U1> Necesito información de profesores.
S2> Por favor, diga el nombre del profesor.
U2> López.
S3> He encontrado dos profesores llamados López: Pedro López y Sara López. ¿A quién se refiere?
U3> A Sara López.
S4> ¿Qué información necesita acerca de la profesora Sara López?
U4> Necesito su número de teléfono.
S5> Perdone, no le he entendido. ¿Qué información necesita?
U5> Teléfono.
S6> El número de teléfono de la profesora Sara López es 958123456. ¿Necesita alguna otra información acerca de los profesores?
U6> No gracias.

Aplicando el esquema automático de anotación propuesto al ejemplo anterrio, se obtienen los siguientes valores de *profundidad* y *anchura*:

U1> Profundidad=1, Anchura=0
U2> Profundidad=2, Anchura=0
U3> Profundidad=3, Anchura=1
U4> Profundidad=4, Anchura=0
U5> Profundidad=5, Anchura=1
U6> Profundidad=6, Anchura=0

Como puede observarse en el ejemplo, el usuario necesitó dos turnos (U2 y U3) para hacer que el sistema comprendiera el nombre de la profesora. En el turno U5 reformuló lo que había dicho en el turno U4, lo que resolvió el malentendido. Por esto *anchura* tiene el valor 1 en estos dos turnos.

El esquema descrito ha sido implementado de forma automática en el sistema empleando la historia del diálogo, que está almacenada en ficheros log. Los valores de la *profundidad* y *anchura* de cada turno de usuario se calculan comprobando el tipo de la intervención previa del sistema. Por ejemplo, como muestra la figura 3.11, la *anchura* sólo sería 0 después de un turno del sistema del tipo "nombre de asignatura" si la intervención actual del sistema fuese "información de asignatura". Si el turno actual fuese "desambiguación de asignatura", la *anchura* se incrementaría en 1 debido a que se necesitaría un turno adicional para aportar el nombre de la asignatura al sistema.

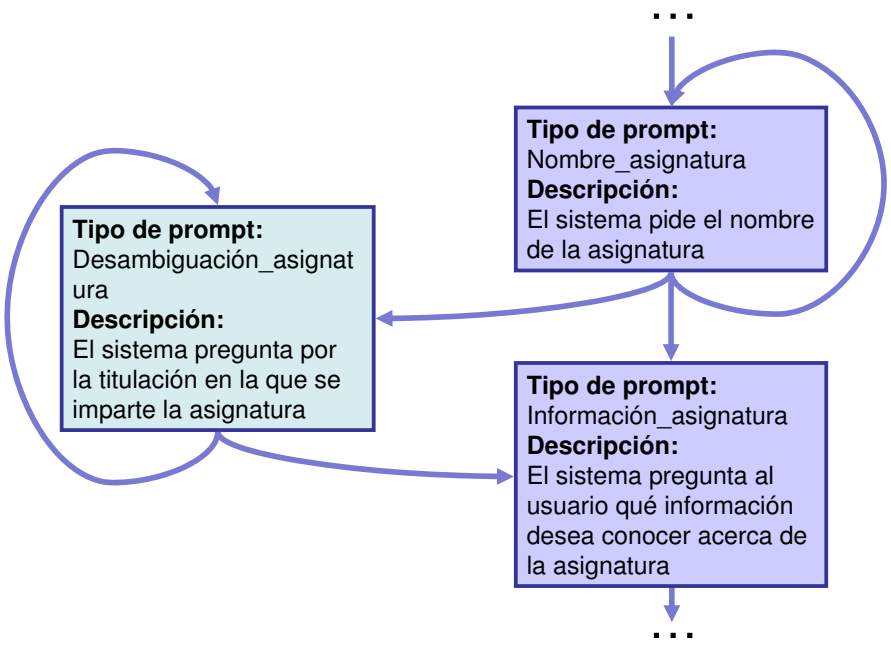


Figura 3.11. *Ejemplo de transiciones entre intervenciones de UAH*

Un estudio exhaustivo del corpus UAH reveló que los diferentes usuarios reaccionan con emociones distintas al mismo estado del diálogo de forma menos predecible de lo que a priori se hubiese podido esperar. Utilizar un umbral para la *profundidad* y otro para la *anchura* para distinguir emociones de forma independiente demostró ser un método ineficiente pues las emocio-

nes se ven influenciadas por una mezcla de ambos parámetros. Además, se obtuvo que no es suficiente con calcular la *anchura* considerando únicamente el turno anterior o el subdiálogo actual, sino que también hay que tener en cuenta la historia completa del diálogo hasta el momento. Este resultado de investigación difiere de las conclusiones a que llegaron Rotaru y Litman (2006), que anotan su variable horizontal ("transición") únicamente con información acerca del turno anterior del sistema. Por ejemplo, en el siguiente diálogo:

(...)
S1> Por favor, diga el nombre del profesor.
U1> Martín.
S2> Perdón, no le he entendido bien. Por favor, repita el nombre del profesor.
U2> Luis Martín.
S3> ¿Ha dicho Luis Marín?
U3> No, Luis Martín.
S4> ¿Qué información necesita del profesor Martín?
U4> Su dirección de correo electrónico.
(...)

la *anchura* sería 0 en U1, 1 en U2, 2 en U3 y 0 de nuevo en U4 porque el diálogo comienza a tratar con un nuevo dato. Un valor alto de la *anchura* en U2 indica que hay una probabilidad alta de que el usuario se enfade debido a los malentendidos y las repeticiones necesarias para hacer al sistema entender el nombre del profesor al que se refiere. Sin embargo, en el turno U4, el usuario puede estar aún enfadado aunque la *anchura* sea 0.

Para solventar esta situación, se ha definido una tercera medida denominada *anchura acumulada*. Mientras que la *anchura* registra el número de turnos adicionales necesarios para que el usuario facilite al sistema un dato en concreto, la *anchura acumulada* denota el número total de turnos extra empleados a lo largo de todo el diálogo. La *anchura acumulada* se inicializa a 0 y se incrementa en 1 cada vez que se incrementa la *anchura*. Por tanto, en el ejemplo anterior, en U3 la *anchura* es 2, lo que indica que fue necesario repetir o confirmar la información acerca del nombre del profesor dos veces. En U4 la *anchura* es 0 de nuevo porque el sistema está recogiendo u nuevo dato, concretamente la información que el usuario desea conocer del profesor. Por tanto, la *anchura acumulada* es más representativa que la *anchura*

porque tiene en cuenta todos los puntos "problemáticos" que se han dado previamente durante el diálogo. Por ejemplo, en U4 la *anchura acumulada* es 2, lo que permite saber que hubo 2 turnos "problemáticos" anteriores al actual.

El algoritmo propuesto en la tesis para clasificar las emociones a partir del contexto del diálogo es el siguiente:

if Alguno de los 2 turnos anteriores se ha anotado como *ENFADO* **then**
  ENFADO
else if $(D \leq 4)$ AND $(A \leq 1)$ **then**
  DUDA
else if $(\frac{A}{D} \geq 0,5)$ OR $((D > 4)$ AND $(\frac{A}{D} < 0,2))$ **then**
  ABURRIMIENTO
else
  ENFADO
end if

donde 'D' denota *profundidad* y 'A' *anchura acumulada*. En el método propuesto, las elocuciones de los usuarios se consideran dubitativas cuando el diálogo ha sido corto y no ha tenido más de un error, pues en las primeras etapas del diálogo es más probable que los usuarios estén inseguros sobre la manera en que deben interactuar con el sistema. Una elocución es reconocida como *aburrimiento* cuando la mayor parte del diálogo se ha empleado en repetir y confirmar información al sistema. Un usuario también puede estar aburrido cuando el número de errores es bajo pero el diálogo ha sido muy largo. Finalmente, a una elocución se le asigna la categoría *enfado* cuando se ha considerado que el usuario estaba enfadado en alguno de los dos turnos previos del diálogo (como se comentó al principio de la sección 3.3.2 con la anotación humana), o cuando la elocución no está en ninguna de la situaciones previas, es decir, que el porcentaje de la longitud total del diálogo compuesto por confirmaciones y/o repeticiones oscile entre el 20 % y el 50 %.

Cuando se considera un baseline que siempre clasifica las elocuciones con la emoción más frecuente, que en nuestro caso es *enfado* (mismo baseline que en la sección 3.4.2), se obtiene una tasa de acierto del 51,67 % al distinguir entre las tres emociones. Este resultado se mejora en un 13,61 % empleando el algoritmo propuesto, que obtiene una precisión del 65,28 %.

3.4.4. Clasificación automática basada en características acústicas normalizadas y contexto del diálogo

Se ha propuesto un método en dos etapas que integra ambas fuentes de información contextual: la forma neutra de hablar del usuario (sección 3.4.2) y el contexto del diálogo (sección 3.4.3). Las características acústicas normalizadas con la forma neutra de hablar de los usuarios se emplean para discriminar si cada frase era de la categoría *enfado* o *dudaOaburrimiento*. Una vez que una elocución se clasificaba como *dudaOaburrimiento*, se emplea la información acerca del contexto del diálogo para distinguir entre *duda* y *aburrimiento*. Además, la información sobre el contexto del diálogo se emplea para clasificar elocuciones como *enfado* si no fueron correctamente clasificadas en el primer paso. Los resultados obtenidos por el método en dos etapas se muestran en la figura 3.12, y son significativos según el t-test.

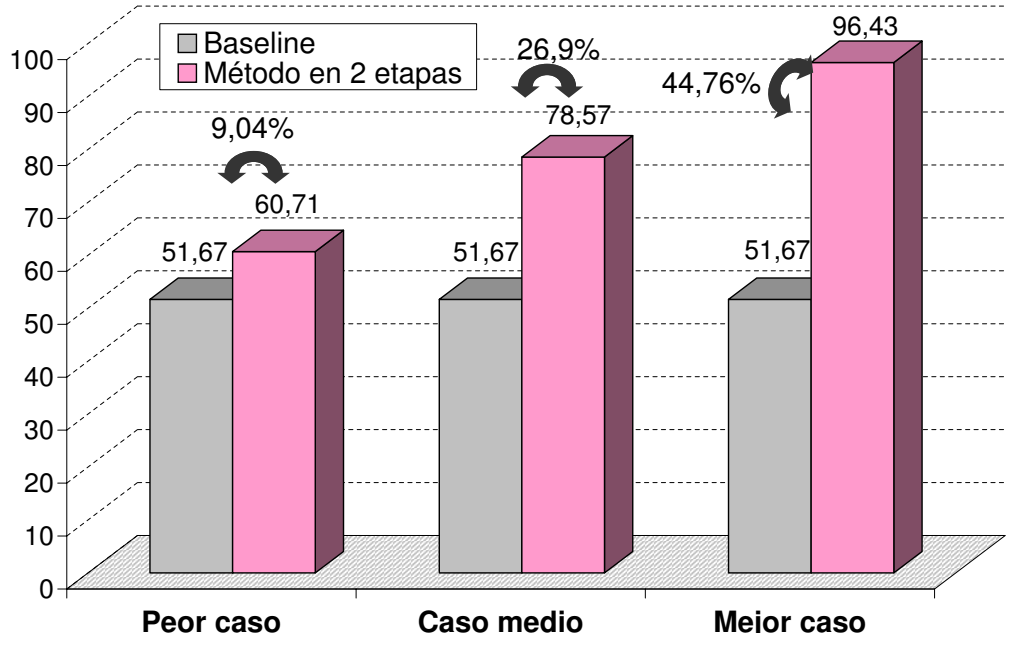


Figura 3.12. *Tasa de reconocimiento de emociones empleando información contextual acústica y del diálogo*

Obviamente, el resultado del método en dos etapas depende del resultado en la primera etapa, dado que el paso de clasificación de *enfado* frente a *dudaOaburrimiento* puede fallar en algún caso y el segundo paso puede tener

que categorizar como *duda* o *aburrimiento* una elocución que pertenezca a la categoría *enfado*. En el peor caso, el primer paso podría fallar en el reconocimiento de todas las elocuciones *enfado*, por lo que todas serían reconocidas como *dudaOaburrimiento* y serían pasadas a la segunda etapa. En este caso, la tasa de acierto es del 60,71 %, pues se ha incorporado un mecanismo para detectar las posibles locuciones de enfado que hayan entrado a la segunda etapa (ver sección 3.4.3). En el caso ideal, el primer paso obtendría una tasa de acierto del 100 % y por tanto clasificaría todas las elocuciones correctamente como *enfado* o *dudaOaburrimiento*. Por lo cual, el segundo paso solamente tendría que clasificar los verdaderamente dubitativos y aburridos. La tasa de reconocimiento en este caso es del 96,46 %. Sin embargo, como se comentó en la sección 3.4.2, con el corpus UAH el primer paso obtiene una precisión máxima del 80,00 %, lo que implica que hasta un 20,00 % de las frases de la categoría *enfado* pueden ser mal reconocidas. Empleando el método en dos etapas, para este caso la tasa de reconocimiento es del 96,43 %. Por tanto, las frases de enfado mal reconocidas pudieron ser correctamente clasificadas en la segunda etapa, obteniendo una tasa de reconocimiento en el mejor caso, en la práctica, idéntica a la del caso ideal.

En media entre el peor y mejor caso, el método en dos etapas obtiene una tasa de acierto del 78,57 % (como puede observarse en la figura 3.13), que mejora los resultados del baseline en un 26,90 %. Dicha mejora es del 44,76 % en el mejor caso, es decir, cuando la primera etapa no comete fallos. La mejora media sobre el reconocimiento basado únicamente en el contexto de la voz neutra es del 9,24 % (27,10 % en el mejor caso). Comparando con el reconocimiento basado únicamente en el contexto del diálogo, la mejora media es del 14,29 % (32,15 % en el mejor caso), y si se compara con los enfoques tradicionales que consideran únicamente características acústicas no normalizadas, la mejora media es del 29,57 % (47,43 % en el mejor caso).

Por consiguiente, emplear únicamente una fuente de contexto (la voz neutra o el contexto del diálogo), mejora tanto el baseline como los enfoques tradicionales donde no se emplea información contextual. Además, combinar ambas fuentes de contexto en el método en dos etapas propuesto, mejora considerablemente al baseline, a los enfoques tradicionales basados en características acústicas sin contexto y a emplear únicamente una fuente de las fuentes propuestas, bien sea el estilo neutro de los usuarios o el contexto del diálogo.

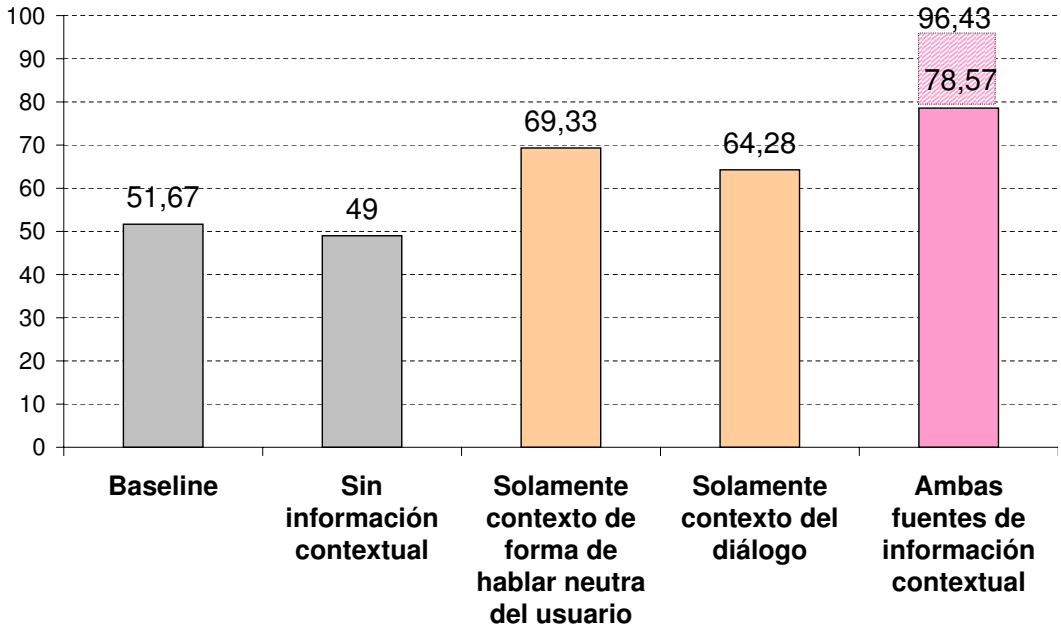


Figura 3.13. *Comparativa de las tasas de acierto de los métodos para reconocimiento automático de emociones*

3.5 Versión previa del método en dos etapas

Antes de encontrar el método óptimo, se realizó un estudio detallado sobre las distintas posibilidades para reconocer las emociones *enfado*, *aburrimiento* y *duda* con un método en dos etapas que empleara las fuentes de contexto propuestas. Los principales objetivos consistían en maximizar: i) la significatividad de los resultados obtenidos, ii) la diferencia entre el baseline y los métodos propuestos en cada etapa, y iii) el empleo de conocimiento contextual en todo el proceso.

Una versión previa del método de las dos etapas distinguía en primer lugar entre *duda* y *enfadoOaburrimiento* empleando el contexto del diálogo, y en segundo lugar entre *enfado* y *aburrimiento* empleando el estilo neutro de hablar de los usuarios, como se muestra en la figura 3.14.

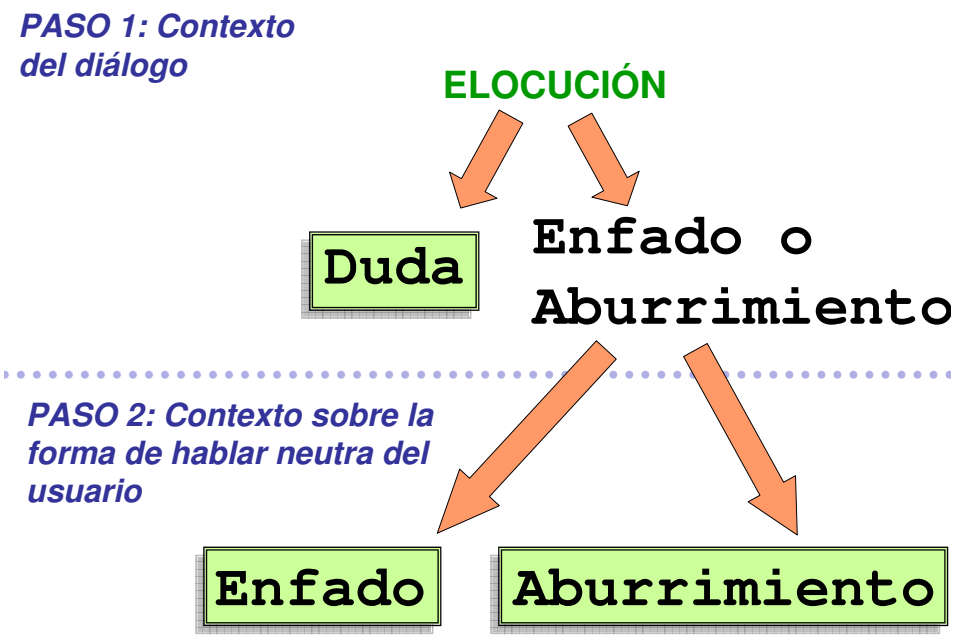


Figura 3.14. *Primera versión del método en dos etapas para el reconocimiento automático de emociones*

En esta primera versión del método en dos etapas, se empleaba un umbral estático T para el contexto del diálogo que se calculaba como sigue:

$$T = D + A$$

donde 'D' denotaba la *profundidad* y 'A' la *anchura acumulada*. En este enfoque, un valor de T mayor o igual al umbral indicaba *enfadoOaburrimiento*, mientras que un valor menor indicaba *duda*.

Se estudiaron diversos valores para el umbral, cuyos resultados se muestran en la figura 3.15. Como puede observarse, para umbrales menores que cinco el porcentaje de elocuciones correctamente clasificadas es menor que el de incorrectamente clasificadas. Para valores del umbral menores, había más correcta que incorrectamente clasificadas. Sin embargo, aunque valores muy pequeños de T como por ejemplo $T = 2$ obtuvieron una mayor diferencia entre el número de instancias correcta e incorrectamente clasificadas, no eran óptimas porque el corpus estaba desequilibrado (había más anotadas como *enfadoOaburrimiento* que como *duda*) y por tanto los mejores resultados se obtenían cuando prácticamente todos las instancias se clasificaban como *enfadoOaburrimiento*. De hecho, las matrices de confusión obtenidas mostraron

que para valores de T menores que cuatro, las elocuciones dubitativas estaban en su mayoría mal clasificadas. Por tanto, se empleó $T = 4$ como valor óptimo para el umbral y el método de clasificación consistió en asignar *duda* a los turnos de usuario con $T < 4$ y *enfadoOaburrimiento* cuando $T \geq 4$; con lo que se obtuvieron tasas de acierto del 70 %.

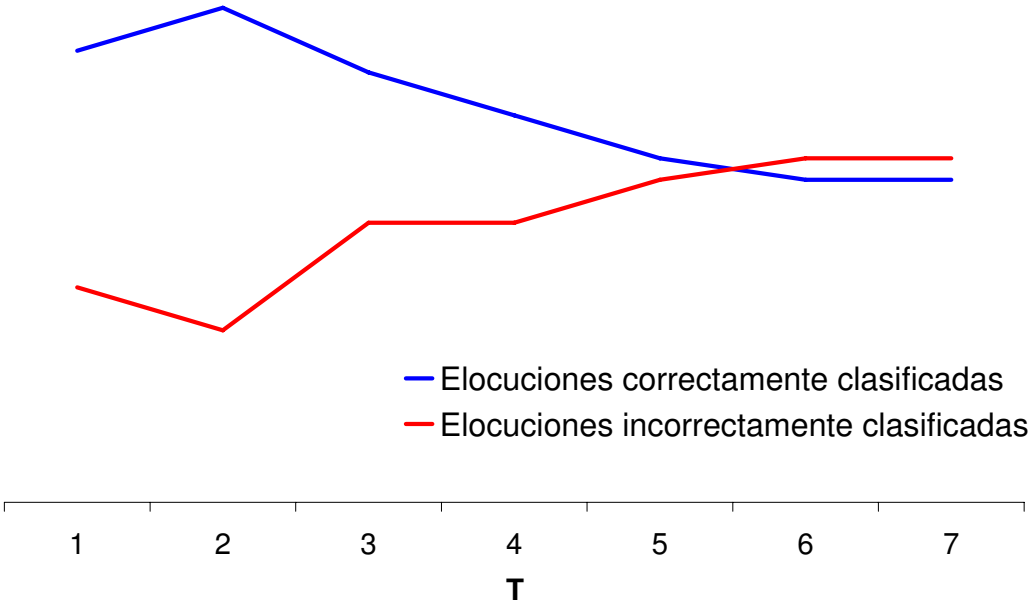


Figura 3.15. *Impacto del valor de los umbrales de contexto del diálogo en el éxito de la clasificación de emociones*

Cuando una elocución era clasificada como *enfadoOaburrimiento*, las características acústicas normalizadas permitían la distinción entre *aburrimiento* y *enfado* siguiendo el mismo procedimiento descrito en la sección 3.4.4. La tasa de clasificación fue de 85,71 % para distinguir entre *enfado* y *aburrimiento*, de esta manera se podía obtener una tasa de clasificación del 60 % para las tres emociones asumiendo que la primera etapa alcanza su funcionamiento óptimo (tasa del 70 %). Por consiguiente, esta propuesta inicial de método en dos etapas supuso una mejora del 24,52 % respecto al caso en que no se utiliza información contextual, sin embargo el resultado era peor que usar las fuentes de contexto por separado, tal y como aperece en la figura 3.16

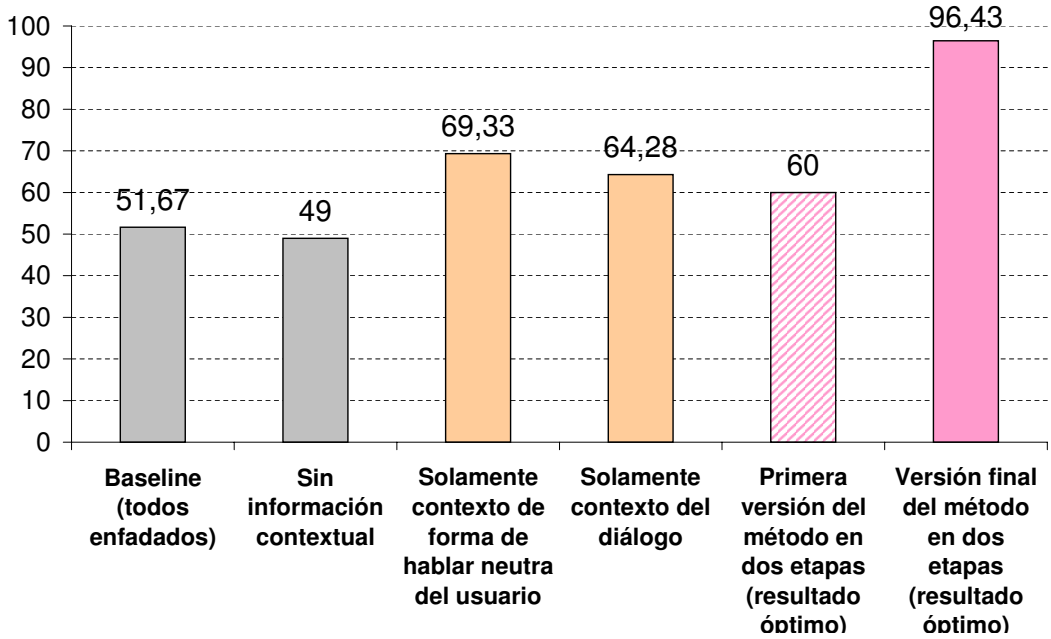


Figura 3.16. *Comparativa de la tasa de acierto de los métodos propuestos para el reconocimiento automático de emociones*

Después de estudiar las características de las emociones a ser reconocidas, se mejoró el método hasta obtener la versión que se ha descrito en la sección 3.4 y que aparece esquematizada en la figura 3.17.

Con este esquema, empleando la información acústica acerca de la voz neutra de los usuarios, se clasificaron las elocuciones como *enfado* o *dudaOaburrimiento*. En la segunda fase, aquellas elocuciones clasificadas como *dudaOaburrimiento* se marcaban como *aburrimiento* o *duda* empleando el contexto del diálogo. Los resultados de la versión final son totalmente comparables con el baseline y el empleo de cada fuente de contexto por separado. Este no era el caso de la primera versión en la que el contexto del diálogo no permitía distinguir entre los enfadados y aburridos, y por tanto no podía ser usada independientemente para clasificar las tres emociones de interés.

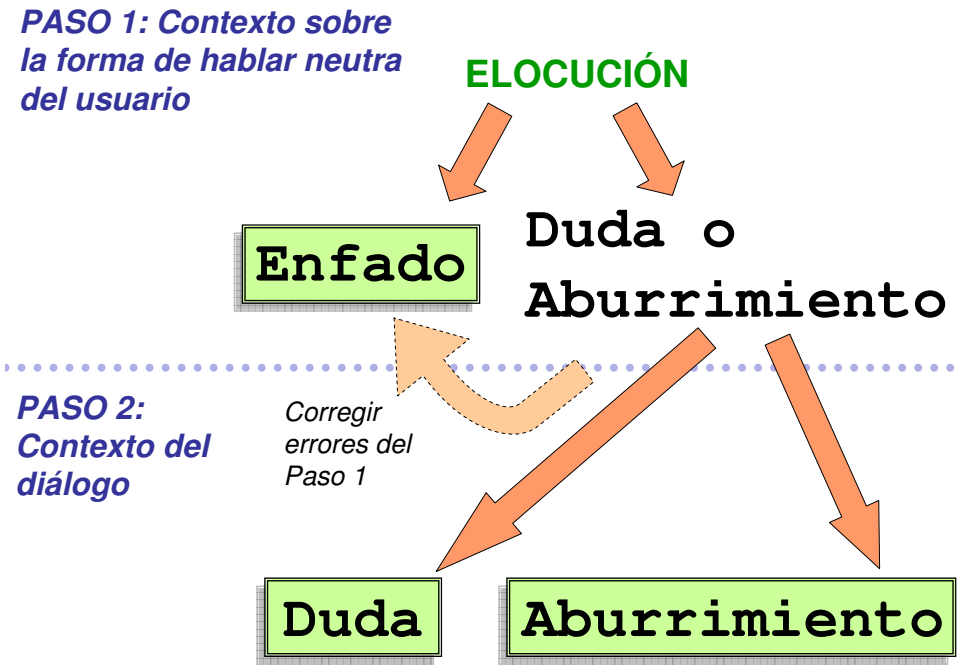


Figura 3.17. *Versión final del método en dos etapas para reconocimiento automático de emociones*

3.6 Conclusiones

En este capítulo se han llevado a cabo diversos experimentos para estudiar la anotación de emociones reales en un corpus obtenido a partir de interacciones de usuarios con el sistema UAH. Los experimentos consideraron tanto anotación manual realizada por nueve anotadores inexpertos, como clasificación automática empleando MLP y distintas técnicas de selección de características. Se ha propuesto el uso de dos tipos de contexto: el de la forma de hablar de los usuarios y el del contexto del diálogo, obteniendo que los anotadores humanos anotaban un 3,40 % más emociones no-neutras cuando hacían uso de conocimiento sobre el contexto. Una razón posible para este resultado es que la información contextual hacía posible identificar habla emocional no trivial (p.ej. detectar emociones presentes de forma más sutil). Por el contrario, cuando se empleaban las características acústicas tradicionales no normalizadas solamente encontraron las emociones más fácilmente distinguibles. Además, se han discutido los problemas que se derivan del uso de corpus de emociones no actuadas en la evaluación de la fiabilidad de las anotaciones. Aunque se ven afectadas de manera irremediable por las deno-

minadas paradojas de los coeficientes kappa, se ha estudiado que el empleo de información contextual durante la anotación ayuda a obtener valores más próximos a las tasas de acuerdo máximas alcanzables, en comparación con los enfoques tradicionales en que este tipo de conocimiento no se emplea.

En cuanto a los métodos de aprendizaje automático, los resultados experimentales muestran que, debido al desequilibrio natural del corpus, es difícil superar el baseline. Esto hace que los métodos tradicionales basados en características acústicas obtengan resultados similares a los de un baseline que siempre etiquete con la emoción más frecuente. Sin embargo, tal y como ocurría con los anotadores humanos, el proceso de clasificación automático mejora sustancialmente al incorporar información sobre la voz neutra de los usuarios y de la historia del diálogo. Simplemente la introducción del contexto acústico de la voz neutra de los usuarios consigue una mejora del 17,66 %. Del mismo modo, emplear información sobre el contexto del diálogo superó al baseline en un 12,67 %. En este capítulo se ha presentado un método en dos etapas para integrar ambos tipos de información contextual. En este método, el contexto referente a la voz neutra del usuario sirve para distinguir entre las categorías *enfado* y *dudaOaburrimiento* con una tasa de acierto del 80,00 %. Si una elocución se clasifica como *dudaOaburrimiento*, el contexto del diálogo se emplea para distinguir entre *duda* y *aburrimiento*. Cuando la primera etapa obtiene una tasa de acierto máxima, el método en dos pasos obtiene una precisión del 96,43 %. En el caso medio, el método propuesto obtiene una tasa de acierto del 78,57 %, que es un 29,57 % mejor que no emplear información contextual, 47,43 % mejor en el caso en que el método tiene un comportamiento óptimo (cuando el primer paso obtiene su tasa de acierto máxima).

Además, los métodos propuestos pueden emplearse durante la ejecución del sistema de diálogo puesto que la información contextual se puede obtener automática y dinámicamente, para ello se ha presentado un método de normalización de las características acústicas respecto al neutro de cada usuario y se ha presentado una representación de la estructura del diálogo basada en dos parámetros que se pueden calcular numéricamente a partir de la información generada por el gestor del diálogo durante la ejecución.

Kolika jazyků znáš,
tolikrát jsi člověkem
Proverbio checo

4

Adaptación de reconocedores de habla a nuevos idiomas

4.1 Introducción

La adaptación entre idiomas (*cross-lingual adaptation*) permite emplear corpus y recursos disponibles en un idioma en el reconocimiento de otro distinto, lo que propicia la implementación de reconocedores del habla de forma rápida y con bajo coste, aunque a costa de un detrimento en la precisión del reconocedor desarrollado. No obstante, esta disminución de precisión puede considerarse insignificante en muchos casos si se compara con el coste que supondría obtener los recursos necesarios para el desarrollo completo de un reconocedor para el nuevo idioma. Por tanto, esta aproximación es especialmente útil para el caso de idiomas minoritarios o dialectos en los que los recursos existentes son muy reducidos o incluso inexistentes.

La hipótesis práctica que se demuestra en este capítulo es que un sistema totalmente funcional para la lengua checa puede adaptarse fácil y rápidamente a su empleo en otros idiomas sin la necesidad de desarrollar nuevos reconocedores y sin requerir un estudio lingüístico arduo, como ocurre frecuentemente en la literatura (p.ej. la inclusión de tildes diacríticas morfológicas presentada por Kirchhoff y Vergyri (2005)). A lo largo del capítulo se propone y estudia una nueva aproximación para alcanzar este objetivo y se muestran resultados experimentales que demuestran su buen funcionamiento tanto para un idioma que es similar al checo (eslovaco) como para un idioma con un origen muy diferente (español).

En primer lugar, mediante la adaptación al idioma eslovaco se quería demostrar la idoneidad de la técnica propuesta para adaptar rápidamente un sistema de reconocimiento de voz existente para su funcionamiento con un lenguaje muy parecido fonéticamente. Además, un objetivo adicional consistía en demostrar que es posible aprovechar la mayor parte del proceso de obtención de todos los recursos necesarios para la creación de sistemas de reconocimiento de voz para un idioma minoritario como el checo (hablado por unos 12 millones de personas), mediante su utilización con un idioma con menos recursos como es el eslovaco (hablado por aproximadamente 6 millones de personas) obteniendo altas tasas de reconocimiento.

En segundo lugar, la principal contribución es la adaptación al español. Su objetivo fue averiguar si esta adaptación entre idiomas también ofrecía buenos resultados con lenguas que pertenecen a familias diferentes y que, por tanto, son fonéticamente distintas. Tal y como puede observarse en la figura 4.1[1], el checo pertenece a la familia de las lenguas eslavas como el ruso, concretamente a la checo-eslovaca junto con el eslovaco. Por otra parte, el español es un idioma itálico como el italiano o el francés, concretamente pertenece al grupo ibero-occidental al igual que el portugués. De este modo, uno de los retos de la tesis fue obtener una correspondencia satisfactoria entre estos idiomas tan distintos (español y checo), teniendo en cuenta que las investigaciones previas habían obtenido resultados muy pobres para tareas de adaptación entre lenguas eslavas e itálicas. Éste es el caso de (Zgank et al., 2004), donde se estudia la adaptación entre el esloveno y el español. Basándose en sus resultados experimentales, los autores recomiendan considerar únicamente idiomas que sean muy similares con el fin de garantizar la máxima superposición entre fonemas. En la literatura es frecuente la utilización de lenguas con las mismas raíces como por ejemplo el italiano y el español (Bonaventura et al., 1997) (familia de idiomas itálicos) o el inglés y el africaans (Nieuwoudt y Botha, 2002) (familia de idiomas germánicos).

[1]Las flechas continuas indican "pertenece a la familia", p.ej. el español pertenece a la familia de las lenguas ibero-occidentales, familia que a su vez pertenece a la ibero-romance. Las flechas punteadas muestran ejemplos de idiomas pertenecientes a otras subfamilias (p.ej. el ruso pertenece a la familia de las lenguas eslavas, en una subfamilia diferente a la eslávico-occidental)

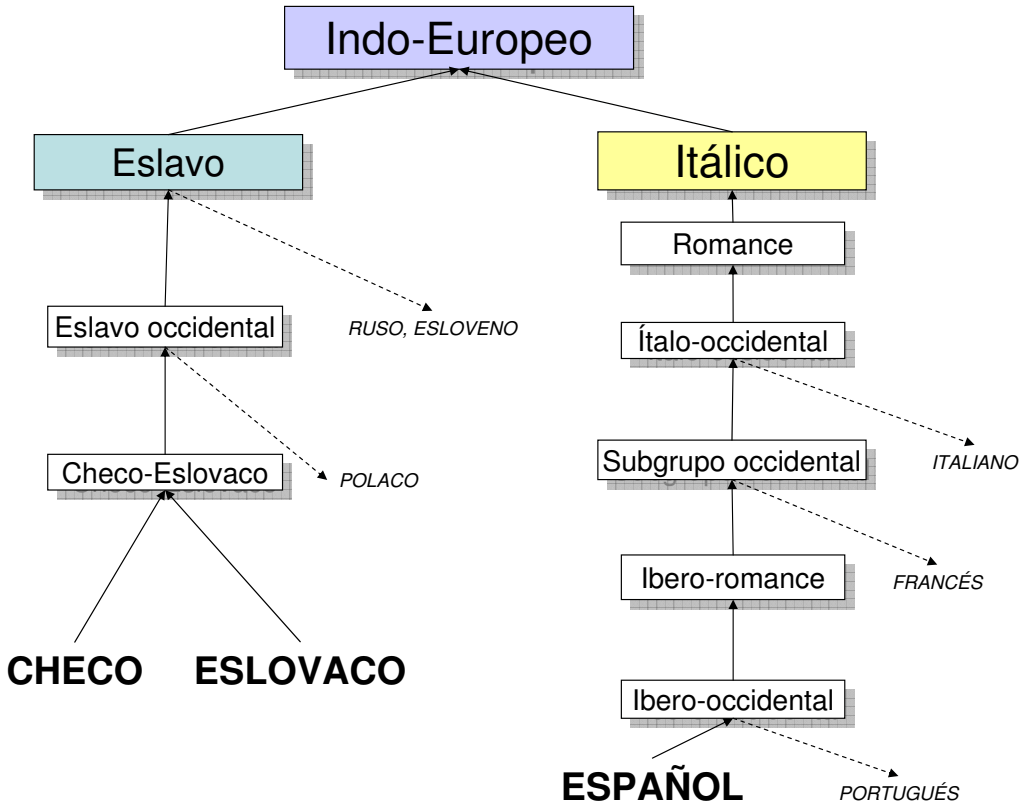


Figura 4.1. *Familias de idiomas del checo, eslovaco y español*

El capítulo está estructurado de la siguiente forma: la sección 4.2 presenta el trabajo relativo a la adaptación entre idiomas, describiendo brevemente la aproximación propuesta en la tesis y comparándola con respecto al resto de metodologías del estado del arte. La sección 4.3 describe el reconocedor del habla checo previamente desarrollado, y que ha sido empleado para el reconocimiento del eslovaco y español, así como el sistema MyVoice, un sistema real en el que se ha evaluado el enfoque propuesto. En la sección 4.4 se detalla la adaptación para cada uno de los idiomas utilizados en la experimentación. La sección 4.5 describe los experimentos que se han llevado a cabo para evaluar el funcionamiento de la técnica propuesta, mientras que la sección 4.6 describe con detalle los resultados obtenidos. Finalmente, la sección 4.7 muestra las conclusiones del trabajo realizado.

4.2 Estado del arte

El término "cross-linguality" se utiliza en muchos contextos dentro del campo de la lingüística computacional, principalmente para describir sistemas que pueden funcionar en varios idiomas. Ésta puede implementarse a diferentes niveles: en el nivel léxico cabe citar las aplicaciones de procesamiento de lenguaje natural como por ejemplo los sistemas de recuperación de texto ("text retrieval"), en los cuales los sistemas de adaptación entre idiomas facilitan la búsqueda y clasificación de documentos escritos en idiomas diferentes a aquel en el que se efectúa la consulta (Fluhr et al., 1999; Martín-Valdivia et al., 2005).

Este término se aplica además a las áreas relativas a la búsqueda de respuestas ("question answering") y el resumen automático de textos ("text summarization") (Radev et al., 2001; Ligozat et al., 2006). Sin embargo, en estas áreas, siempre se requiere una traducción entre los idiomas implicados, que se debe llevar a cabo en algún nivel del procesamiento del texto y la consulta. Del mismo modo, en el dominio semántico se utilizan algunas representaciones intermedias para describir los conceptos y establecer su correspondencia al léxico específico de cada idioma. Por ejemplo, se ha llevado a cabo una investigación recientemente para determinar cómo realizar la correspondencia entre conceptos e imágenes para llevar a cabo búsquedas multilingües en la web (Mihalcea y Leong, 2006).

En el caso de que el lenguaje natural sea oral en lugar de escrito, la adaptación entre idiomas puede aplicarse además a nivel acústico. Tradicionalmente, los sistemas más extendidos que realizan la adaptación acústica entre idiomas son los reconocedores del habla multilingües (Schultz y Kirchhoff, 2006). Éstos se han empleado como componentes de sistemas interactivos basados en la voz como, por ejemplo, sistemas de diálogo multilingües (López-Cózar y Araki, 2005), en los que los usuarios pueden interactuar utilizando diferentes idiomas. Otra muestra son los sistemas de traducción automática orales (*speech-to-speech translation systems*) (Nakamura et al., 2004), que pueden utilizarse como intérpretes en tiempo real. Mediante estos sistemas, un locutor habla utilizando el canal telefónico y su interlocutor recibe la voz traducida en otro idioma.

La adaptación acústica entre idiomas también se ha tratado desde el punto de vista de la compartición de recursos. El desarrollo de un recono-

cedor del habla es una tarea ardua y exigente. Para su implementación, es necesario disponer de una gran cantidad de grabaciones realizadas por centenares de locutores, así como etiquetarlas cuidadosamente para conseguir un conjunto suficientemente representativo para entrenar un modelo acústico. El sistema utilizado en nuestros experimentos para el reconocimiento del idioma checo, descrito en la sección 4.3, se basa en modelos ocultos de Markov de fonemas que se han entrenado con aproximadamente 50 horas de habla anotada proporcionada por 700 hablantes. Para otros idiomas los recursos necesarios pueden ser incluso mayores. Por tanto, la adquisición y etiquetado de los datos precisos requiere generalmente mucho tiempo y esfuerzo. Ésta es la razón por la cual la posibilidad de compartir datos acústicos entre idiomas es muy bien recibida por la comunidad científica. De igual forma, se requiere un gran esfuerzo para desarrollar la parte lingüística de un sistema de reconocimiento del habla. Para el caso del checo fue necesario disponer de un corpus de casi 4 GB de texto, comprendido por 456 millones de tokens con 1,9 palabras diferentes, para obtener un análisis léxico representativo. Ha sido demostrado que para obtener aproximadamente un 99 % de cobertura del checo, debe incluirse un mínimo de 500.000 palabras en un léxico de propósito general. Esto se debe a la naturaleza declinativa del checo, que implica una gran variedad de formas posibles para cada palabra. Sin embargo, el tamaño del vocabulario necesario varía entre idiomas, por ejemplo el inglés requiere únicamente 20.000 palabras para obtener el mismo orden de cobertura (Németh y Zainkó, 2003).

Con frecuencia, estas bases de datos se recopiladas ad hoc por cada equipo investigador y no están abiertas a la comunidad científica. Esto implica que, en la mayoría de los casos, la construcción de un nuevo reconocedor del habla requiere partir desde cero en el proceso de obtención de la totalidad de recursos acústicos y lingüísticos necesarios. Por tanto, los métodos de adaptación entre idiomas son una alternativa a la creación por completo de un nuevo reconocedor. Finalmente, el hecho de compartir recursos acústicos entre idiomas no se ha utilizado en la literatura únicamente para crear un reconocedor del habla a partir de los recursos ya disponibles, sino también para mejorar el funcionamiento de reconocedores del habla utilizando modelos de otro idioma. Por ejemplo, estudios previos han demostrado que el reconocimiento del africaans puede ser mejorado utilizando recursos adicionales en inglés (Nieuwoudt y Botha, 1999).

El uso de la información acústica entre idiomas se ha tratado desde diferentes puntos de vista, que pueden clasificarse en dos aproximaciones principales: aplicaciones multilingües que pueden gestionar múltiples idiomas simultáneamente, y adaptación entre idiomas, donde un reconocedor existente se adapta a una nuevo idioma objetivo.

En la primera aproximación, los reconocedores multilingües son capaces de reconocer simultáneamente varios idiomas compartiendo los modelos acústicos y/o los modelos de lenguaje. Los modelos acústicos multilingües consisten en una colección de modelos acústicos dependientes de cada idioma, o bien, una combinación de modelos acústicos independientes del lenguaje (Schultz y Kirchhoff, 2006). Esta última alternativa se basa en combinar fonemas de varios idiomas en un único modelo acústico. Para determinar qué fonemas de los diversos idiomas se deben combinar en la misma categoría, se han utilizado diversas bases de datos de fonemas multilingües, como por ejemplo GlobalPhone (Schultz y Waibel, 1998). Esta técnica se basa en abstracciones sobre el concepto de *fonema* en unidades de mayor orden tales como *metafonemas* o *archifonemas* (Cahill y Tiberius, 2002), asumiendo que los fonemas de diversos idiomas se pueden agrupar de forma similar a como los alófonos se consideran dentro del concepto de fonema.

La segunda aproximación realiza una adaptación de los reconocedores del habla "monolingües" a otros idiomas. Un posibilidad para realizar esta adaptación consiste en crear correspondencias entre los fonemas. Se han propuesto métodos alternativos como la correspondencia entre palabras (*word mapping*) (Bayeh et al., 2004), sin embargo aunque utilizar este método pueda conducir a veces a mejorar los resultados, es más costoso y menos práctico que si se utilizan fonemas. Esto se debe a que la correspondencia entre palabras es menos propensa a la reutilización, puesto que requiere una traducción completa de todas las palabras posibles junto con sus diversas inflexiones. Por el contrario, los fonemas conforman un conjunto pequeño que puede utilizarse para desarrollar automáticamente cualquier otra construcción de mayor nivel como las palabras.

La idea básica de la correspondencia entre fonemas consiste en establecer una asociación entre las unidades fonéticas de los idiomas origen y objetivo. Así, su resultado depende de la semejanza fonética entre ambos idiomas. Esta correspondencia puede hacerse automáticamente o mediante expertos: el procedimiento automático emplea medidas *data-driven*, que se

extraen con frecuencia de matrices de confusión de fonemas; por otra parte, la aproximación llevada a cabo por expertos se basa en el conocimiento humano sobre los idiomas que se procesan. Generalmente, se utilizan las tablas de símbolos definidas por el International Phonetic Alphabet (IPA) [2] para los diferentes idiomas con la finalidad de determinar los fonemas equivalentes (Schultz y Kirchhoff, 2006), puesto que IPA define una representación única de los fonemas que se pueden utilizar para establecer equivalencias entre idiomas.

En el trabajo que se presenta, se ha utilizado la aproximación de la correspondencia fonética debido a que no se estableció como objetivo construir un reconocedor multilingüe, sino utilizar el reconocedor checo para el reconocimiento de habla en eslovaco o español. Así, el concepto de metafonema no es efectivo para los propósitos definidos. Para realizar la correspondencia, se optó por emplear expertos. Los motivos en los que se basa esta elección son básicamente dos: en primer lugar, aunque la correspondencia automática tenga la ventaja de no requerir la intervención humana y obtener así resultados más objetivos, requiere un número considerable de grabaciones para establecer las semejanzas entre idiomas. Aunque la cantidad de información requerida no sea tanta como la necesaria para construir un nuevo reconocedor completo para el idioma objetivo, este hecho provoca que el proceso de adaptación sea más costoso. En segundo lugar, en el proceso automático los resultados dependen de una correspondencia precisa entre la acústica de ambos idiomas, de la función de distancia definida (Kumar et al., 2005) y de las condiciones de grabación.

4.3 El sistema MyVoice y el reconocedor del habla checo

El reconocedor del habla checo utilizado en la experimentación es fruto del trabajo desarrollado durante más de una década en la Technical University of Liberec (Nouza et al., 2005). Sus modelos acústicos se basan en modelos ocultos de Markov con unidades del habla independientes del contexto que son proporcionadas por varios modelos de ruido, con distribuciones de salida que utilizan al menos 64 gausianas. Dependiendo de las condiciones

[2]Web oficial de IPA: http://www.arts.gla.ac.uk/IPA/

de aplicación, estos modelos pueden ser independientes del locutor(speaker independent, SI), adaptados al género (gender dependent, GD), o adaptados al locutor (speaker adapted, SA). Por otra parte, el módulo de decodificación integrado en el reconocedor utiliza un léxico de palabras ordenadas alfabéticamente, cada una de las cuales se representa mediante su forma escrita y fonética.

Este reconocedor ha sido utilizado satisfactoriamente para el desarrollo de los sistemas MyVoice y MyDictate (Cerva y Nouza, 2007). MyVoice permite que personas con discapacidad motora puedan trabajar con un ordenador empleando varios centenares de comandos orales. Para cumplir este propósito, MyVoice interpreta dichos comandos como una o varias acciones básicas a llevar a cabo; por ejemplo, presionar, mantener o soltar una tecla o combinación de teclas, mover el puntero del ratón y presionar los botones del mismo, ejecutar programas, e imprimir secuencias de caracteres. MyDictate es el primer programa de dictado desarrollado para el idioma checo y dispone de un gran vocabulario que contiene las 540.000 palabras checas más frecuentes.

Para realizar la experimentación presentada en este capítulo, se ha traducido el sistema MyVoice al español y eslovaco. MyVoice se estructura en varios grupos de comandos, cada uno de los cuales se encarga de una tarea específica, por ejemplo el grupo que controla el ratón es diferente del que gestiona el teclado, pero puede accederse fácilmente a todos los demás desde cada uno de ellos usando comandos de voz. El tamaño de estos grupos varía entre 5 y 137 comandos, donde el más grande contiene principalmente los nombres de las letras del alfabeto, y de las teclas del PC, lo que hace que el reconocimiento sea muy difícil debido a que la diferencia acústica entre ellas es muy sutil. Sin embargo, dado que se ha definido un vocabulario específico para cada tarea, se consiguen mejores resultados de reconocimiento cuando se agrupan los comandos. Además, este agrupamiento facilita la interacción, puesto que el usuario es consciente de las palabras válidas que pueden pronunciarse en cualquier momento. El software MyVoice es utilizado en la actualidad por 60 usuarios discapacitados de la República Checa, cuyos informes demuestran que la tasa de palabras erróneas (WER) oscila entre el 1 % y 2 %, para los casos en los que el locutor no posee discapacidades en el habla.

El desarrollo de MyVoice estuvo motivado por la inexistencia de este tipo de herramientas para usuarios checos. Éste es también el caso de muchos otros idiomas para los que las tecnologías del habla no se ha desarrollado suficientemente. El desarrollo de estos idiomas requiere una gran inversión que obtendría pocos beneficios debido a su reducida población. Por tanto, el objetivo fijado para la investigación realizada en la tesis fue la posibilidad de migrar programas como MyVoice a otros idiomas. En primer lugar realizando la adaptación al idioma eslovaco y en segundo lugar, realizando una adaptación más compleja mediante su aplicación al español.

4.4 Adaptación entre idiomas

Para realizar la adaptación entre idiomas se utilizaron textos en eslovaco y en español, conjuntamente con una representación fonética checa generada automáticamente, proceso que se describe en las secciones 4.4.1 y 4.4.2 respectivamente. Los fonemas construidos para el reconocedor checo se aplicaron al reconocimiento de palabras en otras lenguas utilizando la forma fonética checa, y se generaron los modelos acústicos de las palabras concatenando los correspondientes modelos de fonemas. La traducción del texto eslovaco o español a la representación fonética checa se realizó utilizando las correspondencias mostradas en las tablas 4.1 y 4.2 respectivamente. Para realizar la representación fonética se estableció una correspondencia de los símbolos IPA con un conjunto de símbolos ASCII denominado Phonetic Alphabet for Czech (PAC). Existen, otras codificaciones del alfabeto IPA (p.ej. XSAMPA), sin embargo, se ha utilizado PAC por dos motivos principales: en primer lugar, porque se ha establecido como base común para la investigación en el campo del procesamiento del habla en la República Checa (Nouza et al., 1997). En segundo lugar, porque ha sido empleada con éxito por los usuarios de MyVoice, proporcionándoles una forma directa con la cual fijar sus propias transcripciones fonéticas para modificar la pronunciación de los comandos.

Como ilustra la figura 4.2, el resultado de este proceso fue un vocabulario con todas las palabras aceptadas por el reconocedor del habla. Para cada una de ellas, este vocabulario contenía su representación eslovaca o española en modo texto junto con su representación fonética en checo. El reconocedor descrito en la sección 4.3 empleaba este vocabulario como si se tratara de

uno checo, de forma que no fue necesario modificar ni una sola línea de su código. Como resultado, un usuario podía pronunciar una palabra en español o eslovaco, y el reconocedor de voz utilizando los modelos checos obtenía la mejor palabra candidata en español o eslovaco.

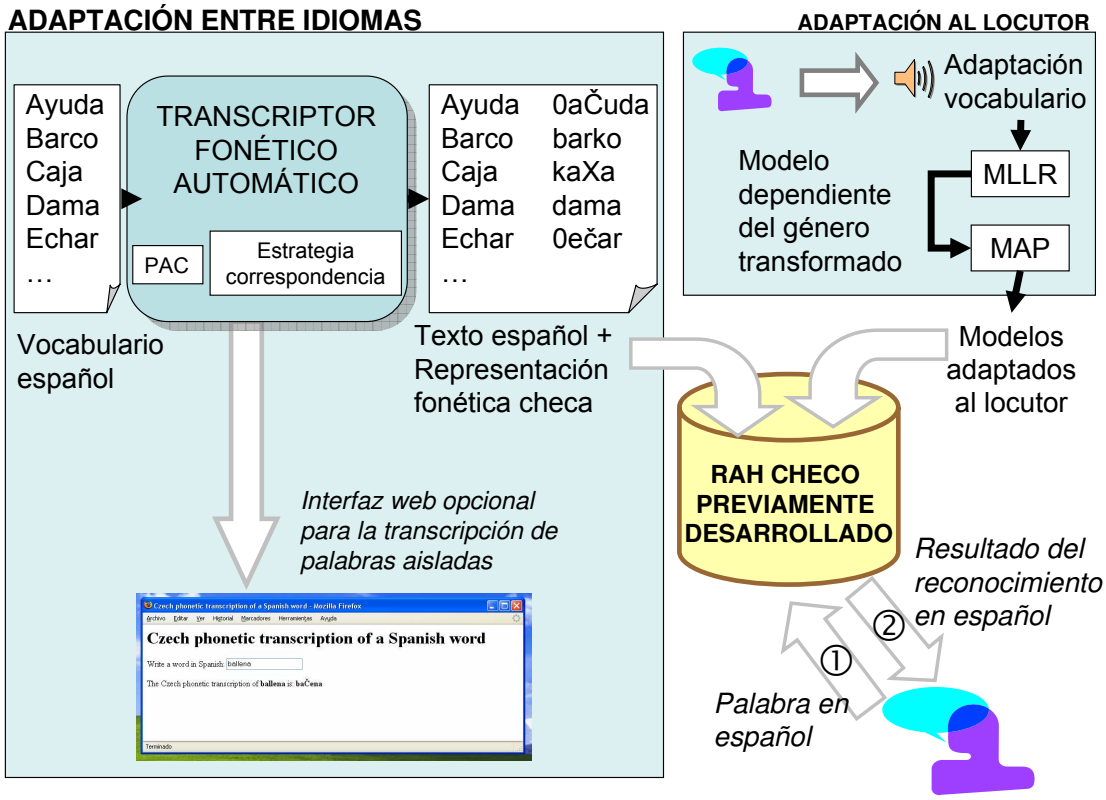


Figura 4.2. *Proceso propuesto para la adaptación entre idiomas*

Para optimizar el funcionamiento del reconocedor, se propone realizar una adaptación al locutor como paso final del proceso. De esta manera, los modelos checos se ajustan para adaptarse mejor a la pronunciación que cada locutor tenga de cada fonema en los diferentes idiomas. Este paso no va en contra del objetivo inicial de implementación con coste eficiente, puesto que puede realizarse de forma rápida y directa haciendo que el usuario lea un texto corto con anterioridad a la primera utilización del reconocedor (p.ej. la primera vez que ejecute MyVoice). La adaptación al locutor se describe en la sección 4.4.3.

4.4.1. Correspondencia fonética entre el checo y el eslovaco

Como se observa en la figura 4.1, el eslovaco y el checo pertenecen a la misma rama de idiomas eslavos, comparten una porción considerable (alrededor del 40 %) de su léxico y muchas de las palabras restantes difieren ligeramente en su deletreo o pronunciación. En general, la lengua eslovaca suena más suave que el checo, debido a pequeñas diferencias fonéticas: el eslovaco utiliza cuatro fonemas adicionales que no están presentes en el checo: 'ĺ', 'l'', 'ŕ' y 'ô'. Tal y como muestra la tabla 4.1, estos fonemas se hicieron corresponder con los fonemas checos más cercanos; en concreto, los fonemas eslovacos representados por los caracteres 'ĺ' y 'l'' con el fonema checo 'l', el fonema eslovaco 'ŕ' con el checo 'r', y el diptongo eslovaco 'ô' con la secuencia 'uo'. El impacto de esta simplificación sobre el funcionamiento del reconocedor es muy pequeño, debido a la semejanza entre el eslovaco y los fonemas checos empleados.

Fonemas en checo	PAC	Fonemas en eslovaco	Ejemplo checo	Ejemplo eslovaco
l	l	l, ĺ, l'	**l**ano	**l**ano, d**ĺ**hý, l'ud
r	r	ŕ	be**r**e	m**ŕ**tvy
uo	uo	ô	d**uo**	m**ô**že

Tabla 4.1. *Correspondencia entre los fonemas eslovacos no presentes en el checo y los fonemas checos más cercanos*

4.4.2. Correspondencia fonética entre el checo y el español

Como describen Zgank et al. (2004), la exactitud de una correspondencia entre idiomas depende del número de fonemas presentes en cada uno y de la semejanza entre los mismos. Es difícil llegar a un consenso acerca del número exacto de fonemas presentes en cada uno de los idiomas utilizado en la experimentación. Esto se da en mayor medida en el caso del español, dado

que posee un número elevado de variedades, incluso considerando únicamente la rama europea y descartando las variantes sudamericanas. Sin embargo, en la literatura, se considera generalmente que el checo dispone de alrededor de 40 fonemas, y el español alrededor de 20, lo que demuestra una gran descompensación entre ambos idiomas.

El resultado de la correspondencia se presenta en la tabla 4.2, que únicamente muestra los fonemas que se emplearon para el reconocimiento del español. Una lista completa de los fonemas checos puede consultarse en (Nouza et al., 1997). Como puede observarse en las dos últimas filas de la tabla, hay dos fonemas españoles que no existen en checo: $/\theta/$ y $/r/$, en la representación de IPA. Para estos fonemas se han estudiado dos soluciones: la primera consistió en utilizar los fonemas checos más cercanos: s y r, en la representación PAC. Esta correspondencia no es artificial, dado que la pronunciación de $\theta/$ como la de $/s/$ está incluso presente en algunas variedades del español, por ejemplo en América latina y algunas áreas de la España meridional. La segunda opción consistió en adaptar estos fonemas checos a la pronunciación española, creando dos nuevos símbolos para la tabla PAC (S y R). Los resultados experimentales fueron muy similares para ambas aproximaciones, con una diferencia de precisión de solamente un 0,5 %, con lo que en adelante la descripción se centrará en los resultados obtenidos empleando la segunda aproximación.

Por otra parte, dado que nuestro sistema no considera alófonos, se ignoraron algunos sonidos, lo que afecta en poca medida a los resultados experimentales obtenidos. Además, existen diferencias en la acentuación que hacen que el reconocimiento del español sea más dificultoso. Las palabras en checo se acentúan siempre en la primera sílaba, mientras que la acentuación en español varía entre palabras. Tal y como se indica en Carreiras et al. (1996), la diferencia en la acentuación en español es importante tanto para el reconocimiento automático del habla como para oyentes humanos. No obstante, el efecto de la acentuación es menos importante en el reconocimiento de palabras aisladas; incluso en los casos de palabras que en español se diferencian por su acentuación, por ejemplo “este” y “esté’, y que no se distinguen al traducirlas al checo.

Fonema en checo	PAC	Fonema en español	Ejemplo checo	Ejemplo español
a	a	a, á	ploch**a**	**a**nillo, **á**guila
b	b	b, v	**b**á**b**a	a**b**uela, **v**ino
č	č	ch	**č**ichá	**ch**arco
ch	X	g, j	**ch**udý	**j**aula, **g**ema
dž	Č	y, ll	rádža	llave, yema
d	d	d	je**d**en	**d**entro
e	e	e, é	l**e**v	**e**so, caf**é**
f	f	f	**f**auna	**f**auna
g	g	g, gu	**g**uma	**g**oma, **gu**isante
i, y	i	i, í	b**i**l, b**y**l	l**i**no, impl**í**cito
k	k	c, k, q	**k**upec	**k**ilo, **qu**eso, **c**asa
l	l	l	de**l**a	**l**ibro
m	m	m	**m**á**m**a	**m**adre
n	n	n	ví**n**o	vi**n**o
ň	ň	ñ	ko**ň**ě	Espa**ñ**a
o	o	o, ó	k**o**l**o**	h**o**la, cami**ó**n
p	p	p	**p**u**p**en	**p**adre
r	r	r	be**r**e	a**r**co
s	s	s	**s**ud	**s**uelo
t	t	t	du**t**ý	**t**eja
u	u	u, ú, ü	d**u**še	l**u**na, **ú**til, ping**ü**ino
-	S / s	c, z	-	**z**umo, **c**ena
-	R / r	rr	-	**r**ueda, pe**rr**o

Tabla 4.2. *Correspondencia entre los fonemas españoles y los más cercanos en checo*

4.4.3. Adaptación al locutor

La aproximación propuesta para la adaptación al locutor es una combinación de los métodos Maximum A Posteriori (MAP) (Gauvain y Lee, 1994) y Maximum Likelihood Linear Regression (MLLR) (Gales y Woodland, 1996), y se realiza en dos pasos. En el primer paso, los vectores de medias

de los modelos checos dependientes del género se transforman utilizando el método MLLR. En el segundo paso, se utilizan estos valores transformados para la adaptación MAP. La ventaja principal de esta aproximación es que los modelos que no aparecen en la adaptación son tratados correctamente por el método MLLR; mientras que MAP asegura que los parámetros de los modelos con muchos datos puedan converger a los valores de un teórico modelo óptimo adaptado al locutor.

Para realizar la adaptación al locutor se utilizó un vocabulario de 614 palabras. Este vocabulario está formado por la lista de las palabras más frecuentes de cada lengua (diseñada para cubrir todos los fonemas), junto con los comandos de MyVoice. Esta decisión se tomó a partir de los resultados experimentales, que mostraron que en la mayoría de los casos, los errores de reconocimiento ocurrieron para palabras cortas (generalmente monosilábicas). Además estas palabras son generalmente las más frecuentes (p.ej. en español son fundamentalmente pronombres, determinantes y preposiciones), por lo que los fallos en su reconocimiento tienen un gran impacto en el cómputo de la precisión del reconocedor. Concretamente, se alcanzó un 44,9 % de mejora relativa para el español al utilizar este vocabulario para la adaptación en lugar de únicamente los 432 comandos de MyVoice. Experimentos adicionales demostraron que esta mejora era mayor que la que se obtiene usando la misma cantidad de fonemas para la adaptación, pero extrayéndolos de palabras seleccionadas aleatoriamente de periódicos en lugar de utilizar las palabras más frecuentes de cada idioma; incluso teniendo en cuenta que ambas listas de palabras consideraban todos los fonemas de cada lengua y éstos se contemplaron en proporciones idénticas.

4.5 Experimentación

Se han llevado a cabo diversos experimentos con el objetivo principal de probar la viabilidad y el funcionamiento del método de adaptación propuesto. Además, se ha evaluado la influencia de varios factores en el funcionamiento del método descrito, como por ejemplo, el uso de diversas estrategias de adaptación al usuario, el tamaño del diccionario de reconocimiento, y el número de palabras consideradas para la experimentación, tal y como muestra la figura 4.3.

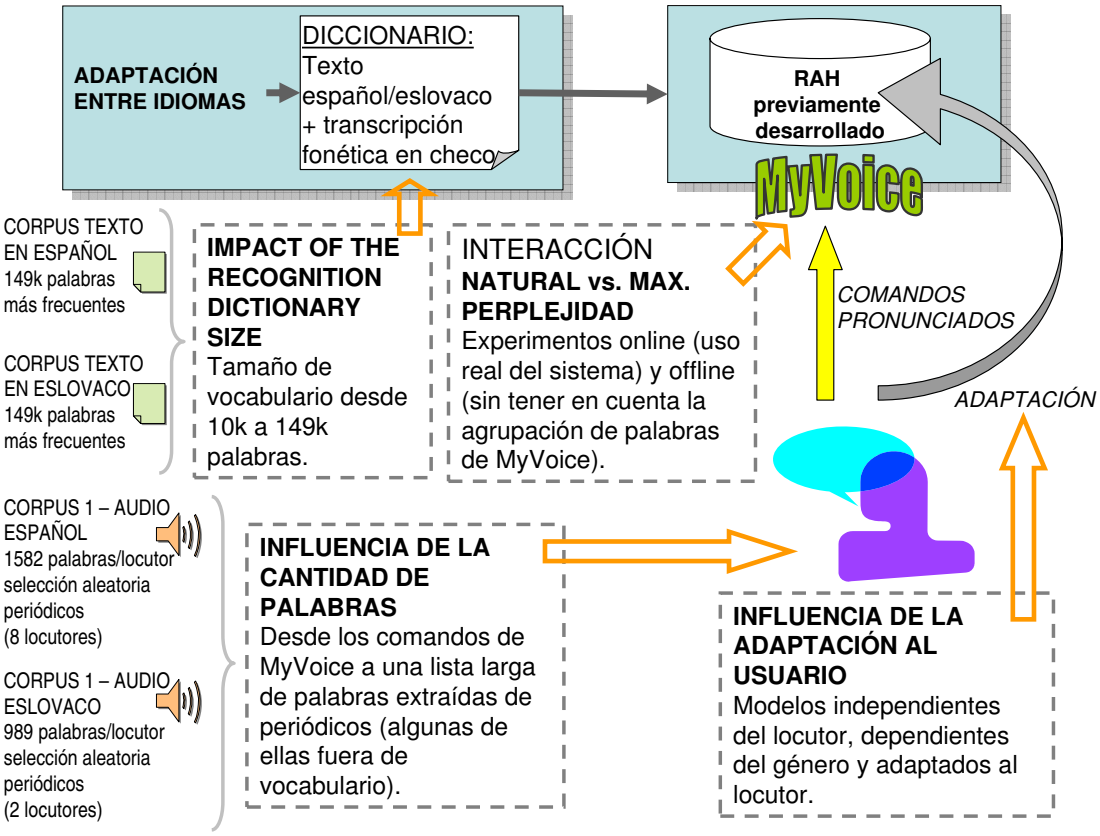


Figura 4.3. *Resumen de la experimentación llevada a cabo*

En primer lugar, se tradujeron los comandos de MyVoice al eslovaco y al español para evaluar el método de adaptación entre idiomas propuesto para una aplicación de órdenes y control (*command-and-control application*). Los usuarios utilizaron MyVoice para controlar su PC mediante comandos orales mientras llevaban a cabo sus actividades diarias, de modo que no se les proporcionó una lista específica de comandos a pronunciar, sino que los resultados se obtuvieron a partir de la interacción natural con el sistema. Tal y como se ha descrito en la sección 4.3, el vocabulario válido de MyVoice se restringe en cada paso a la lista de comandos del grupo actual (vocabulario que puede comprender entre 5 y 137 comandos), siendo el grupo más grande el diseñado para el reconocimiento de nombres de caracteres. Esta tarea supone un gran reto, principalmente debido a la semejanza acústica entre los caracteres; especialmente en algunos conjuntos fácilmente confundibles. Éste

es el caso del conjunto "E-set" del idioma inglés [3] (Odel y Mukerjee, 2007). Por tanto, aunque el agrupamiento de comandos reduzca considerablemente el vocabulario que se debe reconocer en cada instante, el reconocimiento del habla durante la interacción natural con MyVoice no es trivial, puesto que implica la compleja tarea del reconocimiento oral de caracteres.

Con el objetivo de obtener resultados significativos de los diversos modelos de usuario sin importar los grupos utilizados durante la interacción, se llevaron a cabo experimentos adicionales que emplearon el vocabulario completo de MyVoice (432 comandos). En estos experimentos la perplejidad de la tarea fue siempre 432, dado que tras cada comando, podía pronunciarse cualquier palabra.

Además, se quiso corroborar que los resultados obtenidos a partir de la interacción con MyVoice podrían ser alcanzables en situaciones donde el vocabulario aceptado fuese mayor. De este modo, los comandos de MyVoice fueron ampliados con una lista de las 149k palabras españolas y eslovacas más frecuentes, respectivamente. Los dos diccionarios se generaron a partir de periódicos españoles y eslovacos y contienen todas las formas de cada palabra (no sólo los lexemas) que fueron anotadas y ordenadas automáticamente con su frecuencia de aparición.

Para los experimentos se extrajeron subconjuntos de estos diccionarios, cuyo tamaño oscilaba entre las 10k y las 149k palabras, escogiendo siempre las más frecuentes. El objetivo no era construir un sistema eslovaco o español de dictado, sino comprobar simplemente si la aproximación propuesta para la adaptación también funcionaba eficientemente para tareas más complejas con vocabularios cada vez mayores. El tamaño máximo empleado (149k palabras) es mayor que los utilizados generalmente en estudios anteriores, por ejemplo Zgank et al. (2004) varían únicamente el tamaño del vocabulario de dos a varios miles de palabras, mientras que Nieuwoudt y Botha (1999) emplean 60 oraciones como máximo para su evaluación.

Al mismo tiempo que se extendió la lista de comandos de MyVoice, también se aumentó el vocabulario empleado para evaluar el sistema. Para llevar a cabo esta tarea, se seleccionaron aleatoriamente noticias de periódicos españoles y eslovacos, extrayendo palabras de noticias distintas a las utilizadas para generar los diccionarios de reconocimiento, y de categorías

[3] El conjunto "E-set" es un grupo de caracteres del alfabeto inglés fácilmente confundibles. Comprende los caracteres 'B', 'C', 'D', 'E', 'G', 'P', 'T', 'V', 'Z', y a veces además 'M' y 'N'.

diferentes (internacional, economía, arte y deporte). Ocho españoles nativos (cuatro hombres y cuatro mujeres) con edades comprendidas entre 23 y 60 años, y dos eslovacos (un varón y una mujer) con edades de 22 y 24, grabaron las palabras aisladas. Los experimentos con los dos locutores eslovacos usando MyVoice para realizar sus actividades diarias, mostraron un WER casi tan pequeño como para los locutores checos (solamente 2,5 %), siendo el espacio para cualquier mejora muy pequeño. Por tanto, se tomó la decisión de no realizar experimentos adicionales con otros locutores eslovacos, siendo la contribución principal la adaptación al español. Concretamente, cada locutor español grabó 1.582 palabras, mientras que los eslovacos grabaron 989 cada uno. Para evaluar los resultados experimentales se utilizó la media de los valores con respecto al número de palabras para cada idioma, y para todos los locutores. Para obtener resultados que pudiesen ser significativos para el uso del método propuesto en condiciones reales, los corpus no se grabaron en entornos cerrados de laboratorio, sino en el PC de cada locutor, de modo que reflejan las condiciones de ruido verdaderas en las cuales se utiliza el sistema de MyVoice.

Además, para estudiar hasta qué punto la adaptación al locutor permitió que se lograsen mejores resultados de reconocimiento, los experimentos se realizaron con modelos independientes del locutor, dependientes del género, así como adaptados al locutor. Las mejoras alcanzadas por cada uno de estos tres esquemas se compararon considerando el impacto de las palabras fuera de vocabulario (*Out Of Vocabulary - OOV*) y el tamaño del vocabulario empleado para el reconocimiento.

4.6 Resultados experimentales

4.6.1. Interacción con MyVoice

En los experimentos iniciales, los usuarios emplearon MyVoice para controlar su PC mientras realizaban sus actividades diarias. Los experimentos se realizaron de dos formas: on-line y off-line. Los resultados on-line se extrajeron de las interacciones naturales de los usuarios con el sistema MyVoice. De este modo, tal y como se ha comentado anteriormente, la perplejidad de la tarea del reconocimiento variaba de 5 a 137, dado que el vocabulario válido estaba compuesto de las palabras contenidas en el grupo de comandos

actual. En el caso off-line, se utilizaron las mismas frases registradas durante los experimentos on-line, sin embargo el reconocimiento se llevó a cabo empleando como vocabulario válido todos los comandos de MyVoice. Por tanto, como todos los comandos podrían ser pronunciados en cualquier momento, la perplejidad del reconocimiento era 432 (el número total de comandos de MyVoice).

Los resultados experimentales se muestran en la tabla 4.3, como puede observarse, el WER es más bajo para los experimentos on-line, debido a que el tamaño del vocabulario era más pequeño. Al usar modelos adaptados al locutor, se alcanzaron mejoras relativas del 24,1 % para el eslovaco y del 28,3 % para el español con los experimentos on-line y del 46,65 % y 56 % respectivamente para los experimentos off-line. Esto demuestra que la adaptación al locutor redujo considerablemente la diferencia en el funcionamiento para el reconocimiento on-line y off-line, puesto que supone una mejora notable en los resultados off-line, los cuáles son así comparables con los que se obtuvieron en los experimentos on-line (alrededor de un 2 % de WER para el eslovaco y del 4 % para el español).

Idioma	Experimento	Dependiente del género	Adaptado al locutor
Eslovaco	On-line	2,9	2,2
	Off-line	4,6	2,5
Español	On-line	6,0	4,3
	Off-line	10,0	4,4

Tabla 4.3. *WER [en %] para la tarea de comandos y control*

Los experimentos con eslovaco mostraron un WER casi tan pequeño como el obtenido para los locutores checos nativos (solamente 2,5 %). Los resultados con el español fueron mucho mejores de lo esperado a priori, de hecho se diferencian en menos de un 2 % de los que se podrían alcanzar reconociendo la lengua checa.

4.6.2. Efecto de la adaptación al locutor

Los buenos resultados obtenidos utilizando la adaptación al locutor durante la interacción natural con MyVoice, se utilizaron para estudiar hasta

qué punto los modelos adaptados al locutor proporcionan una mejora en la aproximación propuesta. Para probar el funcionamiento de los reconocedores adaptados se utilizaron los corpus en español (12.686 palabras extraídas de periódicos) y eslovaco (1.978 palabras) descritos en la sección 4.5. Además, se utilizó un vocabulario de 10k palabras para el reconocimiento.

Como puede observarse en la figura 4.4, los modelos independientes del locutor posibilitan un WER del 47 % para eslovaco y del 55,8 % para el español. Estos resultados se mejoraron utilizando modelos adaptados al género solamente en un porcentaje relativo del 7,23 % para el eslovaco y del 2,15 % para el español. Sin embargo, la adaptación al locutor posibilitó una mejora relativa del 17,6 % con respecto a los modelos independientes del locutor para el eslovaco, y del 40,8 % para el español. La mayor parte de los errores de reconocimiento para el eslovaco se debieron a las palabras fuera de vocabulario. De este modo, como el objetivo de esta experimentación era medir el impacto de la adaptación al locutor en la aproximación propuesta, sin importar el diccionario y las frases concretas utilizadas para el reconocimiento, se realizaron experimentos adicionales sin considerar las palabras fuera de vocabulario. Según muestra la figura 4.4, el WER disminuyó para ambos idiomas al ignorar estas palabras. En concreto, para el eslovaco el decremento absoluto fue del 20,9 % para el mejor caso. En relación con la adaptación al locutor, se obtuvo una mejora relativa del 54 % con respecto a los modelos independientes del usuario para el español, y del 38,2 % en el caso de eslovaco, obteniendo WERs de alrededor del 20 % para ambos idiomas.

Dado que el eslovaco es un idioma muy similar al checo, el método propuesto logra inicialmente una precisión de alrededor del 70 % (29 % de WER) para este idioma. Por lo tanto, la adaptación al locutor en el eslovaco mejora solamente la precisión en un 11,1 % absoluto (38,2 % relativo) con respecto a usar modelos independientes del locutor. Sin embargo, para un idioma con un origen muy diferente como el español, la adaptación al locutor mejora el reconocedor adaptado notablemente. Los experimentos mostraron un 26,4 % de mejora absoluta (54 % relativa), con tasas de acierto solamente un 4,6 % peores que las obtenidas para el eslovaco. La aproximación propuesta conjuntamente con la adaptación al usuario posibilitó tasas de acierto de alrededor del 80 % (17,9 % y 22,5 % de WER) para el eslovaco y el español.

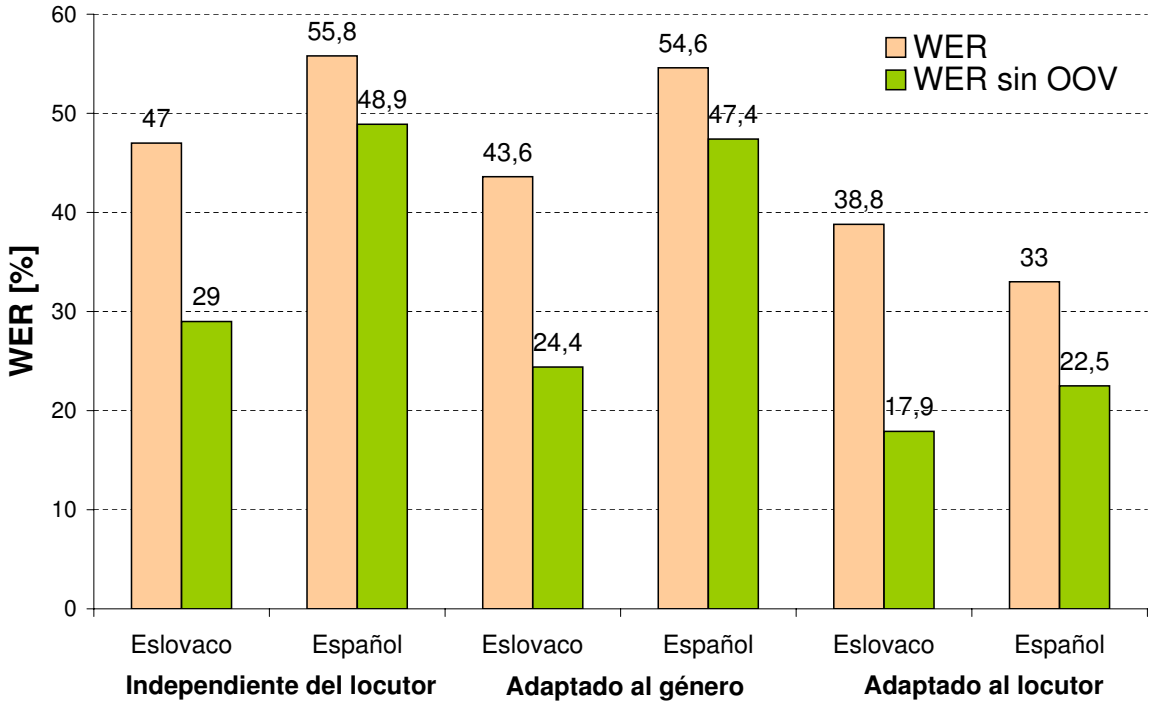


Figura 4.4. *Efecto de la técnica de adaptación en el funcionamiento de los reconocedores adaptados*

4.6.3. Efecto del tamaño del diccionario de reconocimiento

Finalmente, se realizó un estudio para evaluar en qué medida los resultados experimentales podrían variar al aumentar el tamaño del vocabulario del reconocimiento hasta las 149k palabras . Este tamaño es inversamente proporcional al número de palabras fuera de vocabulario que puedan aparecer durante el proceso del reconocimiento, y al mismo tiempo se relaciona directamente con la probabilidad de que una palabra pronunciada sea acústicamente parecida a otra/s del vocabulario.

El número de palabras OOV disminuye drásticamente cuando el vocabulario de reconocimiento es muy grande, por lo tanto el WER tiende a disminuir cuando se utiliza este diccionario. Como puede observarse en la tabla 4.4, la disminución relativa del WER es del 14,7 % para el eslovaco al emplear un vocabulario formado por 149k palabras, en comparación con utilizar 10k, siendo del 15,4 % para el español.

Los resultados experimentales también demostraron que los WER más pequeños se obtuvieron tras la adaptación al locutor, tal y como se observa en

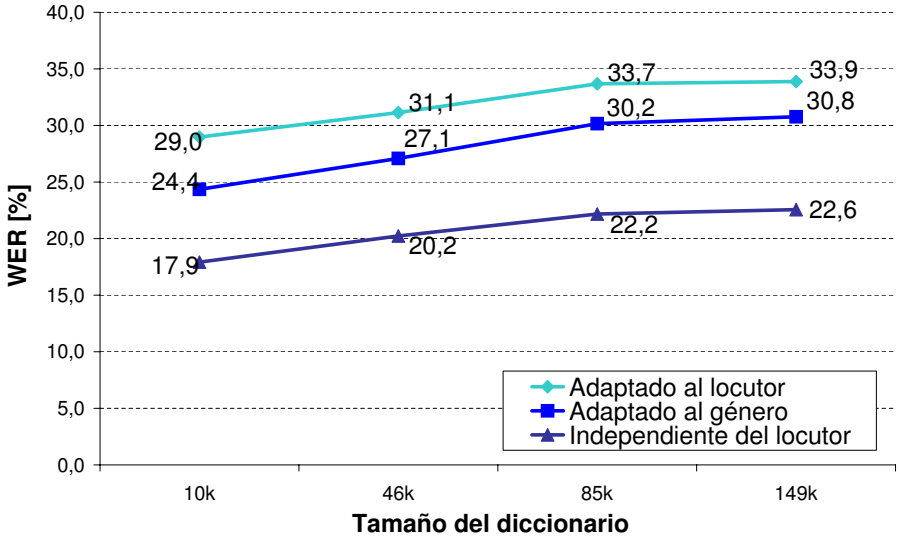
	10k	46k	85k	149k
Eslovaco	38,8	27,3	26,0	24,1
Español	33,0	28,4	28,0	27,9

Tabla 4.4. *Efecto del tamaño del diccionario en el WER [en %] teniendo en cuenta las palabras OOV y la adaptación al locutor*

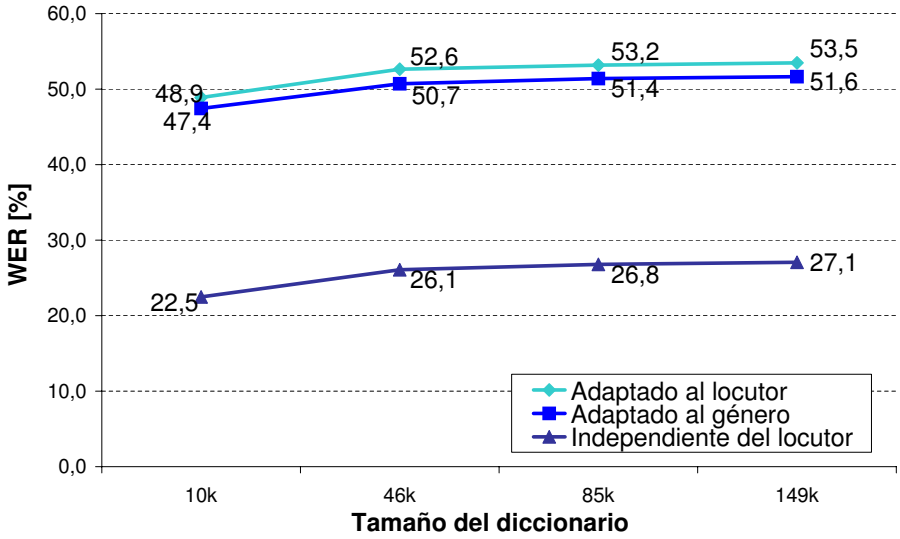
	10k	46k	85k	149k
Eslovaco	38,2	35,0	34,2	33,4
Español	54,0	50,4	49,6	49,4

Tabla 4.5. *Reducción relativa del WER [en %] alcanzada por el reconocimiento adaptado al locutor en comparación con los modelos independientes del locutor*

la figura 4.5. El objetivo era comprobar si otros idiomas se podrían reconocer eficientemente usando el reconocedor checo, sin usar un vocabulario específico para el reconocimiento. Por lo tanto, la figura 4.5 muestra únicamente los resultados obtenidos ignorando las palabras OOV. Como puede observarse, cada paso de adaptación supone una mejora. Esta mejora es más grande para el caso del español, para el que es necesario utilizar modelos adaptados al locutor para obtener resultados similares a los del eslovaco. Según muestra la tabla 4.5, la reducción relativa de WER al pasar de utilizar modelos independientes del locutor a modelos adaptados al mismo, disminuye cuando se acepta un vocabulario más grande. Este hecho se debe a un incremento de la probabilidad de encontrar palabras acústicamente similares en el diccionario. Sin embargo, como puede observarse en la figura 4.5, para diccionarios cada vez más grandes, se obtuvo que el WER tiende a establecerse alrededor del 23 % para el eslovaco y del 27 % para el español. Esto demuestra que la aproximación propuesta, utilizando el procedimiento descrito para la adaptación al locutor, alcanza tasas de precisión de alrededor del 70 % al adaptar del checo al español, es decir dos idiomas con origen muy diverso. Las tasas de exactitud son alrededor del 80 % al adaptar del checo a eslovaco.



(a) Eslovaco



(b) Español

Figura 4.5. *Funcionamiento de los reconocedores adaptados con diversos tamaños de diccionario de reconocimiento y diversas técnicas de adaptación al usuario*

4.7 Conclusiones

En este capítulo se ha descrito la adaptación entre idiomas de un reconocedor del habla checo previamente desarrollado, para el reconocimiento del español y el eslovaco. La adaptación fonética entre idiomas es un área de investigación que está adquiriendo un interés cada vez mayor, debido sobretodo a que permite compartir recursos entre idiomas y representa así una manera factible de desarrollar sistemas para idiomas o dialectos minoritarios. Dado que posibilitar el desarrollo rápido y barato de sistemas basados en el habla es esencial para fomentar su portabilidad, la adaptación entre idiomas se presenta como uno de los desafíos principales en el área (Gao et al., 2005). Sin embargo, los sistemas actuales se basan en estudios lingüísticos y fonéticos que exigen un gran esfuerzo y mucho tiempo. Hasta el momento, se han presentado pocos trabajos que traten el estudio de la adaptación de sistemas de reconocimiento del habla a diferentes idiomas (es decir, de los modelos acústicos, léxicos y lingüísticos) de una manera rentable.

En el presente capítulo se ha demostrado empíricamente que se puede realizar la adaptación de un reconocedor del habla a otro idioma de una manera directa, realizando una correspondencia entre los fonemas, y mejorándola mediante procedimientos de adaptación al idioma y al locutor. Por otra parte, se ha evaluado el método propuesto no sólo con idiomas fonéticamente similares (en este caso el checo y el eslovaco), sino también con idiomas de familias muy distintas, como el checo (eslavo, checo-eslovaco) y el español (itálico, ibero-occidental).

Se han llevado a cabo diversos experimentos utilizando el sistema MyVoice, una aplicación de reconocimiento del habla diseñada para discapacitados checos. El desarrollo de sistemas en nuevos idiomas que presenten una población objetivo tan reducida requiere una inversión que a duras penas puede restituirse. Sin embargo, nuestros resultados experimentales demuestran que para una tarea que implica un vocabulario de 432 comandos, puede lograrse un 95,6 % de precisión (4,4 % WER) para el español y 97,5 (2,5 % WER) para el eslovaco, empleando el procedimiento eficiente en coste que se ha descrito para adaptar un reconocedor checo a dichos idiomas. Además, para vocabularios hasta 149k palabras, el esquema propuesto obtiene alrededor del 72,9 % de exactitud (27,1 % WER) para el español y 77,4 % (22,6 % WER) para el eslovaco.

Ce n'est pas assez de compter les experiences, il les faut poiser et assortir: et les faut avoir digerees et alambiquees, pour en tirer les raisons et conclusions qu'elles portent.

Michel de Montaigne, Essais

5 Evaluación de campo de sistemas de diálogo oral

5.1 Introducción

Como se comentó en la introducción de la tesis, los sistemas de diálogo son cada vez más atractivos para su uso en una amplia variedad de aplicaciones (McTear, 2004; López-Cózar y Araki, 2005; Wahlster, 2006). Con el fin de minimizar costes y optimizar resultados, existe la necesidad de encontrar métodos, arquitecturas y criterios estándar para evaluar, comparar y predecir el rendimiento y la usabilidad de estos sistemas. Desde finales de los años 80 han aparecido diversas iniciativas para establecer dichos métodos. En Estados Unidos la principal institución financiadora para este tipo de investigación es DARPA (Defense Advanced Research Projects Agency), que con su proyecto pionero COMMUNICATOR (Walker et al., 2002a), tuvo como objetivo desarrollar sistemas de diálogo multimodales de forma eficiente. Esto se consiguió empleando diferentes componentes plug-and-play que se evaluaban prestando especial atención a la maximización de la satisfacción del usuario. En Europa, las mayores instituciones relacionadas con la evaluación de los sistemas de diálogo han sido COCOSDA (Coordinating Committee on Speech Databases and Speech I/O Systems Assessment), que se centra en la obtención de corpus que puedan ser compartidos para estudiar criterios de evaluación[1], EAGLES (1996) y DISC (1999). Estos dos últimos proyectos internacionales establecieron un listado de prácticas recomendables para el

[1] http://www.cocosda.org/

desarrollo y evaluación de sistemas de diálogo, tanto a nivel de sistema como de componentes particulares.

Estas investigaciones sentaron las bases para establecer un conjunto común de criterios cuantitativos de evaluación. Sin embargo, no existe un entendimiento ni consenso global acerca de qué criterios deben ser tenidos en cuenta para optimizar la usabilidad de los sistemas de diálogo. Algunos proyectos han intentado abordar el problema de la predicción de la usabilidad y la satisfacción del usuario a partir de criterios de rendimiento medibles. Este es el caso de PARADISE (Walker et al., 2000a), que se ha convertido en uno de los modelos de referencia para la evaluación de sistemas.

Debido a la complejidad y el esfuerzo que demanda la aplicación de este modelo, en la literatura muchos autores aplican medidas cualitativas y cuantitativas por separado. Por ejemplo, Hartikainen et al. (2004) proponen una metodología para la evaluación subjetiva que ha sido empleada para evaluar el sistema multimodal de navegación MUMS (Hurtig, 2004). Recientemente, el sistema VIrtual CO-driver (Geutner et al., 2002), el quiosco multimedia MASK (Lamel et al., 2002) y el sistema de diálogo SAMMIE (Becker et al., 2006), también han sido evaluados únicamente de forma subjetiva. Otros autores como Robinson et al. (2006), evalúan sus sistemas tanto con criterios medidos instrumentalmente como con opiniones de los usuarios acerca de su calidad, pero sin establecer enlaces entre las distintas medidas de evaluación empleadas.

En este capítulo se obtienen resultados empíricos acerca de las relaciones entre ambos tipos de criterios para la evaluación de sistemas de diálogo oral. Esto se llevó a cabo mediante estudios de correlación, que creemos que son el método más fiable que puede aplicarse tanto a la evaluación de sistemas completos como a la evaluación a nivel de componente. Sin embargo, cuando se realizan estudios sobre un gran número de métricas, existe la posibilidad de que algunos de los descubrimientos se deban al azar, y por tanto se han realizado también estudios de fiabilidad y significatividad. Este método se ha empleado de forma exitosa para la evaluación de otros sistemas de diálogo como BoRIS (Möller, 2005), arrojando relaciones muy interesantes entre los criterios de evaluación.

A pesar de ello, los resultados en la literatura están por lo general basados en interacciones restringidas en laboratorio, en las que se pide a los usuarios que interactúen con el sistema siguiendo unos escenarios pre-

establecidos. En algunos casos, también se les proporciona cuestionarios de evaluación en los que expresar su opinión personal acerca de los diferentes aspectos de la interacción. La principal desventaja de este método es que dichos escenarios pueden diferir de las tareas que un usuario habría seleccionado en una interacción no predefinida. Por el contrario, la evaluación de campo se realiza a partir de interacciones de usuarios reales con el sistema final en sus entornos reales. Aunque como señala Bernsen y Dybkjaer (2000), las evaluaciones de campo pueden no ser suficientemente representativas de la funcionalidad completa de los sistemas, creemos que ofrecen los resultados más realistas ya que cubren las verdaderas motivaciones de los usuarios. Las evaluaciones de campo no son repetibles puesto que el contexto de la interacción es muy variable. Esta es también su principal ventaja ya que reúnen resultados de distintos usuarios (distinto sexo, voz, conocimiento, experiencia utilizando el sistema), que emplean distintos dispositivos (teléfonos móviles, teléfonos fijos u ordenadores) y con diferentes entornos (distintas condiciones de ruido). Puesto que los resultados recogidos mediante evaluaciones de campo son robustos a esta heterogeneidad, son más relevantes para la predicción del comportamiento real de los sistemas que los estudios de laboratorio. La contribución de la tesis al estado del arte en la evaluación de sistemas de diálogo oral consiste en la obtención de nuevas evidencias empíricas por medio de un estudio de campo llevado a cabo empleando el sistema UAH.

El capítulo se estructura como sigue. La sección 5.2 presenta una visión general de las principales líneas que se estudian en la literatura. La sección 5.3 describe el cálculo de los criterios de evaluación, distinguiendo entre "parámetros de interacción" (interaction parameters) y "valoraciones de calidad" (quality judgments). La sección 5.4 presenta los estudios estadísticos llevados a cabo, mientras que la sección 5.5 analiza los resultados experimentales. Finalmente, la sección 5.6 presenta las conclusiones obtenidas.

5.2 Estado del arte

La evaluación de los sistemas de diálogo se ha empleado en la literatura para un amplísimo abanico de propósitos, entre ellos, medir el rendimiento de los sistemas, comparar un sistema con sus versiones previas para estudiar la adecuación de los cambios, comparar sistemas distintos y predecir el comportamiento de los mismos.

Con independencia de su propósito, la evaluación se puede llevar a cabo empleando enfoques de "caja blanca" o "caja negra". El primero permite el acceso a los detalles internos del sistema para medir su contribución al rendimiento global. El segundo, trata al sistema como una caja negra de forma que su evaluación está basada únicamente en la respuesta del sistema a las distintas entradas del usuario, independientemente de cómo estén implementados éstos dentro. En ambos casos, la evaluación se puede desarrollar sobre el sistema completo o sobre los componentes individuales. Generalmente, la evaluación se lleva a cabo a nivel de componentes, siendo las áreas de trabajo principales la evaluación del reconocedor del habla, el procesador del lenguaje natural, el gestor del diálogo y el componente de salida oral.

El rendimiento del reconocedor del habla se suele evaluar en términos de medidas generadas automáticamente que calculan el número y la importancia de los errores de reconocimiento. La mayoría de estas medidas tienen un uso generalizado, como por ejemplo el "word error rate" (WER) y el "word accuracy" (WA), que son complementarios en su uso. El WER se define como el número de palabras incorrectamente determinadas (calculado como la sumatoria del número de substituciones, borrados y alteraciones realizados por el reconocedor) dividido por el número total de palabras, y WA se calcula como $1 - WER$. Además de estudiar el correcto funcionamiento del motor de reconocimiento del habla, los enfoques de caja blanca permiten el estudio de componentes internos del reconocedor de voz como por ejemplo los modelos del lenguaje, léxicos o fonéticos. Además, tener información acerca de los componentes del reconocedor puede servir para estudios predictivos del rendimiento final del mismo. Por ejemplo Persia et al. (2007) emplean el rendimiento del algoritmo de separación de fuentes (source separation algorithm) para predecir el éxito de la tarea de reconocimiento del habla en entornos ruidosos. En (Lamel et al., 2000a) se puede encontrar un listado de mejores prácticas para la evaluación de reconocedores del habla.

A pesar del uso generalizado de los criterios, la evaluación de distintos sistemas de reconocimiento difiere substancialmente en las condiciones experimentales, lo que hace muy difícil la comparación de los resultados obtenidos en cada caso. Algunos de estos factores influyentes son las características del vocabulario (p.e. palabras aisladas o habla continua, tamaño del vocabulario o similitud fonética de las palabras), el entorno acústico (p.ej. niveles de ruido), las características de transmisión (p.e. errores de transmisión o niveles

de señal) y las características de los locutores (p.ej. edad, sexo o nivel cultural). Todos estos factores deben tenerse en cuenta al tratar de comparar los resultados de evaluación de distintos reconocedores, lo que hace la tarea muy complicada. Recientemente, algunos estudios se han orientado hacia cómo solventar estas dificultades creando corpus que puedan ser compartidos por los investigadores. El objetivo es obtener un banco de pruebas común para la evaluación y subsecuente comparación de algoritmos de reconocimiento del habla, como es el caso del corpus oral NOIZEUS (Hu y Loizou, 2007).

Para la evaluación de procesadores del lenguaje natural se han empleado medidas similares al WER, podemos encontrar una lista en (Gupta et al., 2006). Una muestra son el "slot error rate" (definido como el número de slots incorrectos dividido por el número de slots actualizados) y el "concept error rate" (CER), que se define como el número de slots incorrectos dividido por el número total de slots rellenos. Se puede encontrar una lista más detallada de métricas de evaluación en (Higashinaka et al., 2004). No obstante, la evaluación de los módulos de procesamiento del lenguaje requiere mayor involucración de expertos que juzguen si los resultados son correctos, y por tanto se necesitan más valoraciones de calidad, uno de los estudios seminales acerca de cómo abordar la evaluación empleando expertos fue desarrollado en el proyecto TSNLP - Test Suites for Natural Language Processing (Lehmann et al., 1996). Por otra parte, evaluar componentes de procesamiento del lenguaje natural es muy dependiente del dominio pues se sustentan en la semántica de la tarea y el detalle hasta el cual se realiza, es decir, el número de conceptos empleados, y por tanto es difícil comparar los resultados de evaluación entre distintos sistemas.

La comparación es incluso más complicada en el caso de los componentes de salida oral, puesto que su evaluación se basa casi por completo en valoraciones de calidad, obteniendo resultados de evaluación muy subjetivos. Las medidas empleadas describen principalmente la calidad y/o naturalidad de la respuesta oral generada, así como la comprensión del mensaje por parte del usuario. Los resultados varían según los usuarios y es por esto que la mayoría de los estudios se centran en sectores de la población específicos, como las personas mayores (Lines y Hone, 2002). En (Gibbon et al., 1997) se puede encontrar una extensa lista de tests de inteligibilidad, prosodia y calidad global para evaluar salidas de voz sintetizada.

La gestión del diálogo se suele evaluar en términos de la calidad de la interacción entre usuario y sistema: adecuación de las respuestas del sistema, estrategias de retroalimentación, duración del diálogo, número de turnos medio, adecuación de la iniciativa, adecuación de las estrategias de confirmación o habilidad para solventar situaciones de error. El grupo de trabajo DISC (1999) propone seis grandes grupos de criterios que deben tenerse en cuenta para evaluar gestores de diálogo: gestión correcta de la información acerca del contexto actual del diálogo, correspondencia entre las unidades semánticamente significativas en la entrada más reciente del usuario, contribución específica del usuario, generación de respuesta a la salida al usuario, cuestiones específicas de la evaluación del gestor del diálogo (p.ej. estrategias de retroalimentación) y cuestiones globales de la evaluación de sistemas de diálogo (p.ej. tiempo en completar la tarea).

Es difícil separar la evaluación del gestor del diálogo del resto del sistema, especialmente en los valoraciones de calidad, es por esto que algunos autores proponen corregir las entradas al gestor para ser capaces de tener un "gold standard" para representar el comportamiento del gestor del diálogo de forma aislada. Este baseline puede emplearse sucesivamente para la comparación con situaciones reales en las que el gestor del diálogo se ve afectado por errores en los módulos de reconocimiento y comprensión (Roque et al., 2006a). Por otra parte, el gestor del diálogo hace un uso intensivo de conocimiento acerca del dominio y por tanto debe ser capaz de construir y adaptar el contexto de interacción. Por ello, algunos autores como (Hanna et al., 2007) han estudiado el modo de evaluar gestores de diálogo en términos de la modificación, mantenimiento y reusabilidad del conocimiento del contexto y la gestión del discurso.

No existe consenso en la literatura acerca de la terminología a emplear para categorizar los criterios de evaluación descritos. Tradicionalmente, los autores han diferenciado entre criterios de evaluación objetivos y subjetivos. Los primeros engloban a las medidas que pueden ser calculadas a partir del rendimiento del sistema como el WER, mientras que las segundas consideran aquellas medidas que juzgan alguna propiedad como es la inteligibilidad de la respuesta. Esta notación se ha empleado en la mayor parte de los estudios anteriores, como en Larsen (2003), Minker et al. (2004b) y Robinson et al. (2006). Sin embargo, como argumenta Möller (2005), siempre hay sujetos involucrados en la determinación del rendimiento del sistema. Así, en las de-

nominadas medidas objetivas también se emplean evaluadores humanos con frecuencia, como es el caso del cálculo del WER para el que los expertos tienen que comparar la entrada real del usuario con la salida del reconocedor. Por tanto, Möller (2005) propone diferenciar entre "valoraciones de calidad" (subjetivo), "parámetros de interacción" (que pueden ser tanto medidos instrumentalmente como por expertos) y "predicciones de calidad" (medidas instrumentalmente). La tesis se centra en las dos primeras categorías.

Se ha tratado en varias ocasiones de crear una lista exhaustiva de criterios de evaluación empleando parámetros de interacción, predicciones de calidad y valoraciones de calidad. Así lo hacen Dybkjaer y Bernsen (2000), quienes proponen una lista de 15 criterios para garantizar la usabilidad de los sistemas: uso adecuado de las modalidades, precisión en el reconocimiento de la entrada, flexibilidad en el vocabulario aceptado, calidad de la voz del sistema, adecuación de la generación de respuesta, cobertura adecuada del dominio y satisfacción del usuario, entre otros. El Expert Advisory Group on Language Engineering Standards (EAGLES, 1996), propuso medidas cuantitativas (p.ej. el tiempo de respuesta del sistema) y cualitativas (p.e. la satisfacción del usuario), que se aplicaron e interpretaron siguiendo un marco de trabajo innovador. Este marco proporcionó directrices para llevar a cabo la evaluación y para garantizar la disponibilidad de los resultados de forma fácilmente interpretable y comparable. En el proyecto DISC (1999) se establecieron pautas adicionales que completaban la propuesta de EAGLES empleando metodologías de desarrollo de ciclo de vida.

Otros autores se han centrado en la obtención y estudio de corpus orales para calcular medidas de evaluación. Se trata de corpus largos extraídos del empleo del sistema o de diálogos entre seres humanos. En este último caso, el comportamiento humano puede emplearse como baseline para comparar con el comportamiento del sistema (Paek, 2001). De este modo, el proyecto EVALDA (Devillers et al., 2004) se orienta a campañas de evaluación que consideran diversos aspectos de la interacción en lenguaje natural. Una de ellas es la campaña MEDIA, que evalúa la interacción entre usuarios y sistemas de diálogo. Su metodología de evaluación emplea baterías de tests obtenidas a partir de corpus reales junto con criterios de evaluación de uso extendido. Degerstedt y Jönsson (2006) desarrollaron la herramienta LINTEST para llevar a cabo la evaluación de los sistemas de diálogo empleando el corpus JUNIT. En (Dybkjaer et al., 2004; López-Cózar y Araki, 2005) se

puede encontrar una revisión detallada acerca de los esfuerzos internacionales para la generalización de criterios de evaluación, mientras que Möller et al. (2007) presenta una revisión de los criterios de-facto extraídos de todos estos estudios y un ejemplo de su empleo para evaluar un sistema de diálogo concreto.

Como se ha comentado anteriormente, PARADISE (Walker et al., 1998b) es el método de evaluación más ampliamente aceptado y empleado propuesto hasta la fecha para especificar la contribución relativa de varios factores al rendimiento global del sistema. Este método modela el rendimiento como una función ponderada de: tasa de acierto, eficiencia del diálogo (duración, turnos del sistema, turnos del usuario, número total de turnos), calidad del diálogo (WA, latencia de la respuesta) y satisfacción del usuario (rendimiento del sistema texto-a-voz, facilidad de la tarea, pericia del usuario, comportamiento esperado, uso futuro). En otras aplicaciones, PARADISE ha sido empleado para desarrollar modelos de predicción de la satisfacción del usuario, de nuevo basados en una combinación lineal ponderada de distintas medidas (Walker et al., 2000b). El propósito de este método de evaluación consiste en maximizar la satisfacción de usuario maximizando el éxito de la tarea y minimizando los costes de la interacción tal y como se muestra en la figura 5.1. Estos costes se cuantifican empleando distintas medidas de eficiencia y calidad, donde el peso de cada medida se calcula a través de una regresión lineal multivariable considerando la satisfacción del usuario como variable dependiente y el éxito de la tarea y las medidas de eficiencia y calidad como variables independientes. Este método se ha mejorado recientemente para realizar la evaluación de sistemas de diálogo multimodales. Por ejemplo, ha sido empleado en el proyecto SmartKom, para el que se creó el marco de trabajo PROMISE (Beringer et al., 2002).

La aplicación de PARADISE a la evaluación de sistemas de diálogo requiere corpus de diálogo extraídos de experimentos controlados en los que los sujetos tienen que evaluar su satisfacción en una escala después de haber interactuado con el sistema. Este enfoque ha sido empleado con éxito para evaluar y comparar ocho sistemas COMMUNICATOR (Walker et al., 2002b,a), primero con experimentos de laboratorio controlados, y en segundo lugar en un contexto menos restringido donde los sistemas estaban accesibles a través del teléfono. Estrictamente, esta segunda evaluación no fue un estudio de campo pues los autores tenían control sobre los usuarios, que fueron

reclutados específicamente y asignados a distintos sistemas. En cualquier caso, las tareas que tenían que completar no estaban siempre predefinidas. Un enfoque parecido se empleó en el proyecto ARISE (den Os et al., 1999), donde la evaluación se basaba en respuestas de sujetos que o bien llamaban al sistema de diálogo desde casa, o bien interactuaban con él en el laboratorio. En ambas situaciones, las tareas que los usuarios debían llevar a cabo estaban predefinidas (Sanderman et al., 1998).

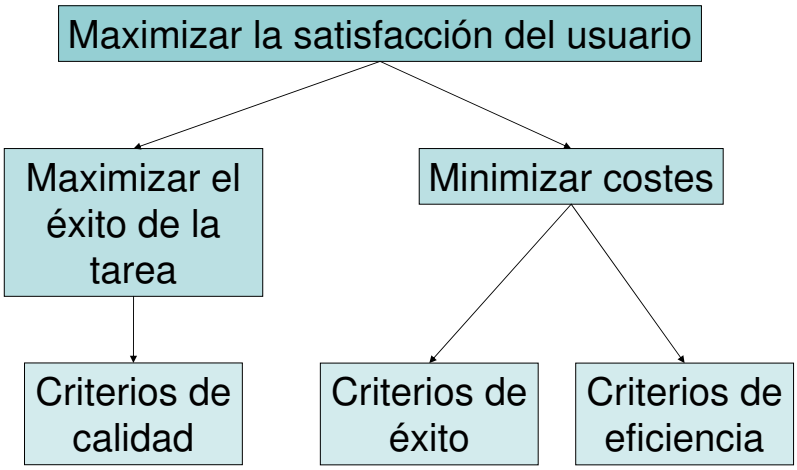


Figura 5.1. *Arquitectura del modelo PARADISE*

También puede encontrarse en la literatura una distinción entre tests "internos" y "externos" según sean llevados a cabo por usuarios pertenecientes al equipo de desarrollo del sistema de diálogo (evaluación interna) o por usuarios que no tengan conocimiento previo acerca del sistema (evaluación externa). Sin embargo esta notación no es equivalente a la distinción entre estudios de campo y de laboratorio, pues los tests externos pueden involucrar el uso de escenarios predefinidos. Así, Rajman et al. (2004) proponen una metodología para el prototipado rápido de diálogos que produce, para cualquier aplicación, una interfaz de diálogo de implementación rápida que puede ser mejorada a través de un proceso iterativo de Mago-de-Oz. Para refinar los modelos de diálogo desarrollados con esta metodología, los autores proponen emplear un test interno y otro externo. El test interno se utiliza para adaptar el prototipo y sus sucesivas modificaciones, mientras que el externo se emplea para la evaluación final de la interfaz de diálogo resultante. En ambos casos la evaluación se lleva a cabo con una encuesta de satisfacción que se les facilita a los usuarios tras haber interactuado con el prototipo siguiendo

un conjunto de escenarios predefinidos que involucran contextos específicos para una búsqueda de restaurantes.

Para estudiar las implicaciones de las evaluaciones de campo, algunos autores se han centrado en evaluaciones no restringidas. Este es el caso del sistema Let's Go (Raux et al., 2003), que fue evaluado empleando interacciones de usuarios reales que llamaban al sistema para conseguir información acerca de horarios de autobuses. La evaluación se llevó a cabo mediante parámetros de interacción (Raux et al., 2006). Desafortunadamente, aunque estos parámetros son relativamente sencillos de calcular, no dan suficiente información acerca de la calidad de la interacción. Las valoraciones de calidad, por el contrario, son difíciles de obtener y comparar al estar relacionadas con opiniones subjetivas. Solamente en ciertos casos los parámetros de rendimiento que pueden ser medidos cuantitativametne son capaces de expresar también calidad.

Nuestro trabajo se orienta hacia el empleo de medidas estándar defacto tanto cualitativas como cuantitativas (Möller et al., 2007) en un estudio de campo con el sistema UAH (capítulo 2). El objetivo principal para la tesis es obtener empíricamente relaciones entre estos parámetros mediante el empleo de estudios estadísticos. En el área de la aceptabilidad de sistemas se han realizado estudios parecidos, específicamente para predecir la adopción de nuevas tecnologías (p.ej. en estudios de riesgo de compañías inversoras). Uno de los modelos más empleados es el Technology Acceptance Model, que relaciona criterios de usuario con la adopción final de las tecnologías por éstos (Legris et al., 2003). No obstante, no se consideran parámetros cuantitativos en este modelo. En el área de los sistemas de diálogo, muy pocos autores han empleado estudios de correlaciones para medir dichas relaciones, así lo han hecho Litman y Pan (2002), Möller (2005) y Schiel (2006) los han aplicado a estudios de laboratorio.

5.3 Criterios de evaluación

La evaluación del sistema UAH se llevó a cabo utilizando tanto parámetros de interacción como valoraciones de calidad. Los parámetros de interacción se han empleado para medir el rendimiento del sistema (p.ej. número de errores cometidos por el reconocedor del habla), y el curso del diálogo (p.ej. duración del diálogo o número de turnos). Estas medidas per-

mitieron llevar a cabo diversos estudios sobre el rendimiento y fiabilidad del sistema así como descubrir puntos de interacción que pudieran ser mejorados. Aunque los parámetros de interacción fueron un buen indicador de la calidad de la interacción evaluada, no aportan necesariamente información acerca de la satisfacción del usuario (López-Cózar y Araki, 2005). Por tanto, es necesario llevar a cabo una evaluación cualitativa registrando las opiniones de los usuarios en relación con estos aspectos de la interacción. En los experimentos presentados en este capítulo, la evaluación subjetiva se llevó a cabo utilizando cuestionarios.

5.3.1. Parámetros de interacción

Para calcular los valores de los parámetros de interacción, se ha empleado el corpus UAH como se comentó en el capítulo 2, este corpus está compuesto por 85 diálogos y 422 turnos de usuario, con una media de 5 turnos por diálogo. Cada diálogo se anotó automáticamente con dos sellos de tiempo correspondientes al momento de inicio y fin de la interacción. Dentro de cada diálogo, las elocuciones de los usuarios se almacenaron en ficheros .wav junto con información sobre su hora de comienzo, el turno previo del sistema y los resultados de reconocimiento asociados a la misma, es decir, las palabras reconocidas y los niveles de confianza en el reconocimiento asociados a las mismas.

Después se registró manualmente si cada elocución había sido entendida correctamente por el sistema con independencia de los errores de reconocimiento en las palabras. De esta manera, si en respuesta a la intervención del sistema: "¿Qué tipo de información desea?", el usuario responde: "Quiero información acerca de una asignatura", pero la frase reconocida fue: "Información acerca de asignaturas", hubo dos borrados y dos alteraciones. Sin embargo, independientemente de estos errores, la elocución fue entendida correctamente por el sistema, pues los valores semánticos devueltos por la gramática de reconocimiento fueron correctos. Por tanto, esta elocución se registraría como "correctamente comprendida".

A nivel de diálogo, se anotaron el sexo del usuario, si el diálogo se completó (es decir, si el usuario no colgó antes de que acabase) y si la interacción tuvo éxito. Al ser un estudio de campo, no había tareas predefinidas que los usuarios debieran completar, por tanto se tuvo que definir una es-

trategia para considerar el éxito del diálogo. Concretamente, se estableció marcar los diálogos como exitosos cuando los usuarios hubieran conseguido la información que habían requerido.

Todas las anotaciones se almacenaron en una base de datos a partir de la que se calculaban de manera automática los valores de los parámetros de interacción. Así, la duración del diálogo se obtuvo a partir de los sellos de tiempo y el número de turnos de confirmación empleando la información acerca del turno previo del sistema. La tabla 5.1 expone los parámetros de interacción empleados en los experimentos y la figura 5.2 muestra cómo se realiza el cálculo de los parámetros de interacción con un diálogo de ejemplo.

Parámetro	Descripción	Necesidad de anotación manual
Éxito de la tarea	Valor binario que indica si el usuario obtuvo la información que pidió al sistema	Sí
Completitud del diálogo	Valor binario que indica si el usuario esperó hasta el final del diálogo para colgar	Sí
Duración del diálogo	Duración del diálogo en segundos	No
Número de turnos de usuario	Número de turnos de usuario en el diálogo	No
Media de palabras por turno en el diálogo	Número medio de palabras en todas las elocuciones del diálogo	Sí
WER	Número de palabras mal reconocidas dividido entre el número total de palabras pronunciadas por el usuario durante el diálogo	Sí
Confianza de reconocimiento media	Confianza media de las palabras en todos los resultados de reconocimiento del diálogo	No
Tanto por ciento de elocuciones correctamente comprendidas	Porcentage de elocuciones correctamente comprendidas con respecto al número total de elocuciones del diálogo	Sí
Número de turnos de confirmación	Número de veces que el sistema pidió explícitamente confirmación durante el diálogo	No

Tabla 5.1. *Parámetros de interacción empleados*

DIÁLOGO

S1> Bienvenido al sistema UAH, ¿en qué puedo ayudarle?

U1> Querría información de profesores.

S2> Por favor diga el nombre del profesor.

U2> Juan Tena.

S3> ¿Ha dicho Juan Tema?

U3> No, Juan Tena.

S4> ¿Qué información desea acerca de D. Juan Tena?

U4> Su horario de tutoría.

S5> El horario de tutoría del profesor Tena es lunes y martes de 10 a 12. ¿Desea más información acerca de los profesores?

U5> No, gracias.

S6> Gracias por utilizar el sistema UAH. Que tenga un buen día.

FASE 1 :
Grabación de llamadas y almacenamiento automático de información

2006_04_07, 15:02:48 – 15:03:59, Masculino

	Marca de tiempo	Turno previo del sistema	Frase reconocida	Confianza del reconocedor
U1	15:03:09	Menú_inicial	Quiero información de los profesores	0,89
U2	15:03:20	Preguntar_nombre_prof	Juan Tema	0,4
U3	15:03:36	Confirmar_nombre_prof	No Juan Tena	0,9
U4	15:03:43	Menú_profesores	Su horario de tutoría	0,85
U5	15:03:57	Más_información	No gracias	1

FASE 2 : Cálculo de los parámetros de interacción

2.1: Anotación por un experto

Éxito del diálogo	1
Completitud del diálogo	1
Número de palabras por turno	U1 – 4 U2 – 2 U3 – 3 U4 – 4 U5 – 2
Número de inserciones, borrados y alteraciones por turno	U1 – 1 alteración, 1 inserción U2 – 1 alteración
Correctitud de la interpretación semántica	U1 – 1 U2 – 0 U3 – 1 U4 – 1 U5 – 1

2.2: Cálculo automático

Dialogue duration	71
Number of user turns	5
Average recognition confidence	0,81
Number of confirmation turns	1
Average number of words/turn	3
WER	0,2
%correctly understood utterances	8%

Figura 5.2. *Ejemplo de cálculo de los parámetros de interacción*

5.3.2. Valoraciones de calidad

La interacción con UAH comienza con un mensaje de bienvenida en el que el sistema se presenta y le pide al usuario que visite una web donde puede completar un cuestionario con su opinión sobre el rendimiento del sistema. Para poder unir estos resultados con las grabaciones de la interacción correspondiente, se da al usuario un número de identificación del diálogo que posteriormente se le pide en el cuestionario junto con la fecha en que realizó la llamada y una hora aproximada de comienzo de la interacción.

El cuestionario es el siguiente:

Q1. Puntúe de 1 a 5 su conocimiento de las nuevas tecnologías de acceso a la información (1="Bajo", 5="Alto").
Q2. Puntúe de 1 a 5 su uso previo de sistemas automáticos de diálogo telefónico (1="Bajo", 5="Alto").
Q3. ¿Cuántas veces había utilizado el sistema UAH con anterioridad?

- No lo había usado antes.
- ..... veces.

Q4. ¿Cómo le entendía el sistema a usted?

- Muy mal.
- Mal.
- Aceptablemente.
- Bien.
- Muy bien.

Q5. ¿Cómo entendía usted los mensajes que generaba el sistema?

- Muy mal.
- Mal.
- Aceptablemente.
- Bien.
- Muy bien.

Q6. La conversación le ha parecido:

- Muy lenta.
- Lenta.
- Adecuada.
- Rápida.
- Muy rápida.

Q7. Corregir los errores que quizás haya cometido el sistema le ha parecido:

- Muy difícil.
- Difícil.
- Fácil.
- Muy fácil.
- El sistema no ha cometido errores.

Q8. ¿Ha sido fácil averiguar la información que necesitaba conocer?

- No, ha sido totalmente imposible.
- Sí, pero con gran dificultad.
- Sí, pero con dificultad.
- Sí, ha sido fácil.
- Sí, ha sido muy fácil.

Q9. ¿Está satisfecho/a con el funcionamiento del sistema?

- No, nada.
- Casi nada.
- Indiferente.
- Satisfecho/a.
- Muy satisfecho/a.

Q10. ¿Tenía claro lo que debía hacer en cada momento del diálogo?

- No, nunca.
- Casi nunca.
- A medias.
- Casi siempre.
- Sí, siempre.

Q11. ¿Cree que el sistema se ha comportado de forma "similar" a como lo haría un ser humano en esta tarea?

- Nunca.
- Casi nunca.
- A medias.
- Casi siempre.
- Siempre.

Las respuestas a cada pregunta se codificaron y almacenaron convenientemente en la base de datos. A todas las respuestas excepto las correspondientes a Q3 se les asignó un valor entre uno y cinco (en el mismo orden en que aparecen en el cuestionario). Los valores por defecto fueron: Q1=1, Q2=1, Q3=1, Q4=3, Q5=3, Q6=3, Q7=5, Q8=3, Q9=3, Q10=3 y Q11=3. A partir de los resultados del test, se extrajeron las medidas enumeradas en la tabla 5.2.

Parámetro	Pregunta de la que se ha extraído
Conocimiento acerca de las nuevas tecnologías de acceso a la información	Q1
Conocimiento acerca de los sistemas de diálogo	Q2
Experiencia usando el sistema UAH	Q3
Percepción de hasta qué punto UAH entiende al usuario	Q4
Percepción de hasta qué punto el usuario entiende a UAH	Q5
Velocidad de interacción percibida	Q6
Presencia percibida de errores cometidos por UAH	Q7
Facilidad percibida de corregir los errores de UAH	Q7
Facilidad percibida de conseguir la información requerida	Q8
Satisfacción del usuario	Q9
Hasta qué punto el usuario sabía qué hacer en cada momento de la interacción	Q10
Percepción de comportamiento similar al humano de UAH	Q11

Tabla 5.2. *Parámetros de calidad percibida y perfil de los usuarios*

Las tres primeras medidas que aparecen en la tabla 5.2 no son valoraciones de calidad, sino información referente a los usuarios. Con la ayuda de estas preguntas, se pretendió obtener una idea aproximada de su perfil.

Puesto que los usuarios de UAH son principalmente estudiantes y profesores de nuestra escuela, el conocimiento sobre las nuevas tecnologías de acceso a la información era bastante alto en todos los casos, tal y como muestra la figura 5.3. Sólo el 36 % de los participantes del test fueron mujeres.

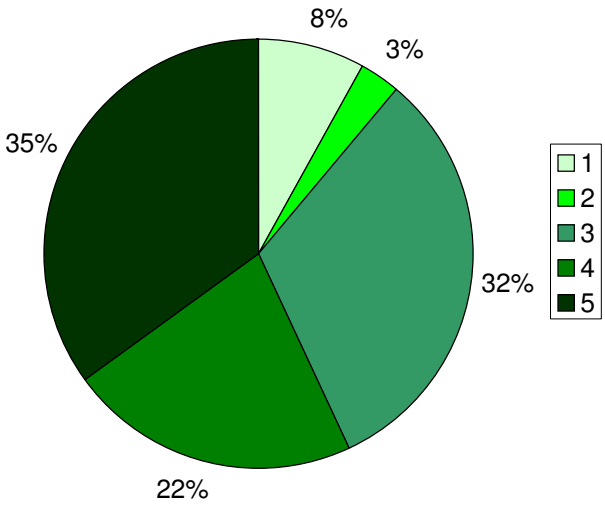


Figura 5.3. *Conocimiento de los usuarios acerca de las nuevas tecnologías de acceso a la información (1 = Bajo, 5 = Alto)*

Como los experimentos se basan en llamadas de usuarios realizadas por iniciativa propia, creemos que los resultados son realistas dado que la interacción surge de una necesidad real de los usuarios. Además, los diálogos fueron más heterogéneos, puesto que tuvieron lugar en distintos contextos. La desventaja de este enfoque fue que, aunque se animó a los usuarios a responder los cuestionarios, algunos no lo hicieron y por tanto no hay registradas valoraciones de calidad para todos los diálogos grabados. Concretamente, sólo 37 de los 85 diálogos cuentan con medidas subjetivas además de las objetivas. La figura 5.4 muestra los datos demográficos de ambos tipos de usuarios: los que respondieron el test subjetivo y los que no.

Como puede observarse, de los diálogos correspondientes a los usuarios que no rellenaron el test subjetivo, en el 8,47 % el sexo del usuario era desconocido. Esto se debe a que los usuarios colgaron después de la primera intervención del sistema y antes de responder a la misma. La primera frase de UAH dejaba claro que se trataba de un sistema automático y que la llamada iba a ser grabada con fines de investigación. Por tanto, creemos que hay dos razones posibles por las que los usuarios colgaran antes del primer turno:

bien que no se sintieran cómodos hablando con un ordenador, o bien que no estuvieran de acuerdo con que se grabaran sus interacciones.

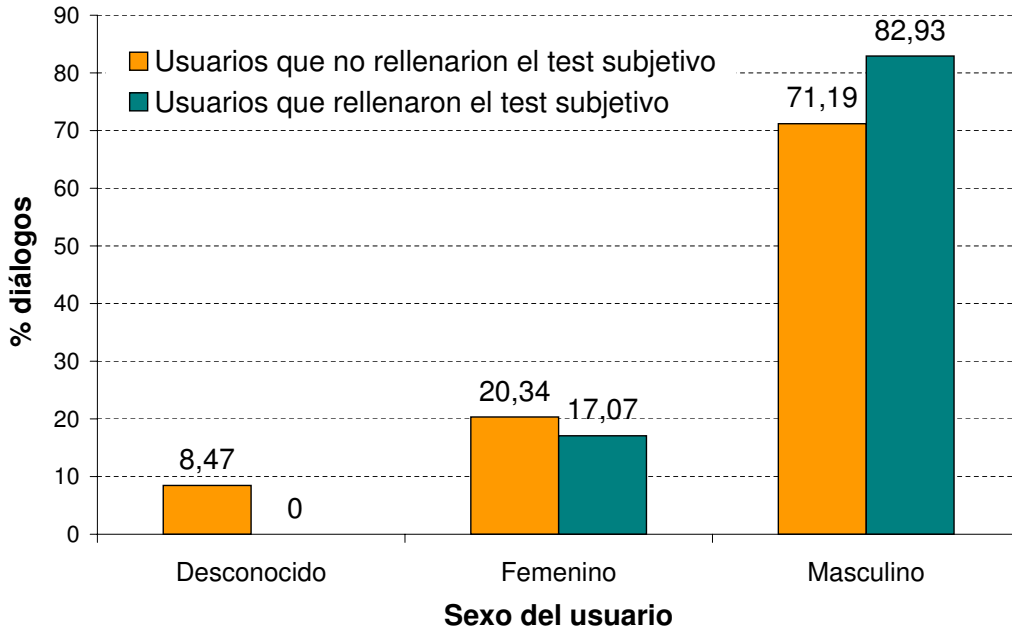


Figura 5.4. *Datos demográficos para los distintos tipos de usuario*

Los estadísticos descriptivos de todos los parámetros según los tipos de usuarios se muestran en la tabla 5.3, donde se indican los valores mínimo y máximo, así como el rango de todas las medidas empleadas. La sección 5.5.1 presenta un estudio detallado de las diferencias en rendimiento y calidad percibida entre las interacciones de ambos grupos.

5.4 Estudios estadísticos utilizados para la evaluación

Para encontrar relaciones relevantes entre los criterios utilizados, se correlacionaron todas las variables, obteniendo el valor absoluto del *coeficiente de correlación de Pearson*. Sin embargo, el valor del coeficiente de correlación por sí mismo no es bastante para obtener resultados fiables, siendo también necesario calcular la probabilidad de obtener los resultados por azar. Con este fin se calculó la significatividad (o *p-value*) de cada coeficiente

Parámetro	Tipo de usuario	Rango	Min,	Máx,	Media	Desv, Típica	Varianza
Conocimiento acerca de las nuevas tecnologías de acceso a la información	Test	4	1	5	3,77	1,14	1,30
Conocimiento acerca de los sistemas de diálogo	Test	4	1	5	3,23	1,28	1,65
Experiencia usando el sistema UAH	Test	9	1	10	2,80	3,20	10,22
Percepción de hasta qué punto UAH entiende al usuario	Test	4	1	5	3,69	1,25	1,57
Percepción de hasta qué punto el usuario entiende a UAH	Test	2	3	5	4,37	0,69	0,48
Velocidad de interacción percibida	Test	3	1	4	2,71	0,62	0,39
Presencia percibida de errores cometidos por UAH	Test	1	0	1	0,54	0,50	0,25
Facilidad percibida de corregir los errores de UAH	Test	3	1	4	2,47	0,90	0,82
Facilidad percibida de conseguir la información requerida	Test	4	1	5	3,37	1,437	2,06
Satisfacción del usuario	Test	4	1	5	3,63	1,09	1,18
Hasta qué punto el usuario sabía qué hacer en cada momento de la interacción	Test	3	2	5	4,29	0,893	0,798
Percepción de comportamiento similar al humano de UAH	Test	4	1	5	3,57	1,04	1,08
Éxito de la tarea	Test	1	0	1	0,77	0,43	0,18
	Sin test	1	0	1	0,46	0,50	0,25
Completitud del diálogo	Test	1	0	1	0,74	0,44	0,20
	Sin test	1	0	1	0,36	0,48	0,23
Duración del diálogo	Test	153	21	174	96,66	37,06	1373,70
	Sin test	297	0	297	90,14	64,65	4179,88
Número de turnos de usuario	Test	9	1	10	5,34	2,26	5,11
	Sin test	16	1	17	4,7	3,94	15,52
Media de palabras por turno	Test	3	1	4	1,81	0,69	0,48
	Sin test	4,33	0	4,33	1,73	0,78	0,61
WER	Test	0,67	0,00	0,67	0,19	0,18	0,03
	Sin test	0,83	0	0,83	0,25	0,28	0,05
Confianza de reconocimiento media	Test	0,16	0,82	0,98	0,93	0,04	0,002
	Sin test	0,23	0,77	1	0,93	0,05	0,003
% elocuciones correctamente comprendidas	Test	0,50	0,50	1	0,95	0,12	0,15
	Sin test	0,74	0,33	1	0,89	0,19	0,04
Número de turnos de confirmación	Test	2	0	2	0,80	0,63	0,40
	Sin test	3	0	3	0,62	0,88	0,77

Tabla 5.3. *Estadísticos descriptivos de los criterios usados*

de correlación. Si el nivel de significación es muy pequeño (menos de 0,05) entonces la correlación se entiende significativa y se considera que los dos criterios están relacionados linealmente.

Dado que la mayoría de las variables estaban intercorrelacionadas, se ha estudiado el efecto que cada criterio ejercía en la significatividad de las relaciones entre los demás. Es posible que dos criterios estén correlacionados únicamente por verse influenciados por un tercero (eliminando el efecto de éste no estarían correlacionados). Para estudiar las relaciones aisladamente, eliminando el efecto del resto de los criterios, se midieron los *coeficientes de correlación parcial* conjuntamente con sus niveles de significación.

El *coeficiente de correlación de Pearson* funciona correctamente para las variables escalables, cuyos valores representan categorías ordenadas siguiendo métricas significativas, como la duración del diálogo en segundos, de modo que sea apropiado realizar comparaciones de distancia entre valores. Sin embargo, para la investigación descrita no se utilizaron únicamente variables escalables, sino también variables ordinales y dicotómicas (se muestra una clasificación en la tabla 5.4). Los valores de las variables ordinales representan categorías con una graduación intrínseca, como en el caso de los parámetros de calidad percibida descritos en la sección 5.3.2, mientras que las variables dicotómicas, como el "éxito de la tarea" la "completitud del diálogo", poseen únicamente dos valores (0 o 1 en nuestro caso). Para obtener resultados fiables, se generaron tablas de contingencia para los criterios ordinales, con las que es posible estudiar estas variables y descubrir asociaciones entre ellas. Para medir la magnitud de sus relaciones, se utilizaron los coeficientes *Tau-b de Kendall* y *Rho de Spearman*. La interpretación de estos coeficientes es equivalente a la del *coeficiente de Pearson*, sin embargo, como se basan en características ordinales de los datos, sus valores y significatividades pueden no coincidir.

Además, se llevaron a cabo análisis de varianza (ANOVA), que intentan describir una variable dependiente como resultado de la suma ponderada de varios factores. Concretamente, se utilizó el test *one-way ANOVA*, en el cual existe únicamente una variable independiente, y se calculó el *coeficiente F*. Cuando el valor crítico de F está por debajo de 0,05, es posible descartar la igualdad entre las medias y concluir que no todas las medias poblacionales que se están comparando son iguales. Se calculó además *Eta square*, que es una valoración del grado en el cual afecta cada factor a la variable dependiente. Para obtener más información en la que basar las interpretaciones realizadas, y especialmente para el caso de las variables dicotómicas, también se calculó la *V de Cramer*, que permite contrastar la hipótesis de independencia en las tablas de contingencia.

Parámetro	Tipo
Conocimiento acerca de nuevas tecnologías de acceso a la información	Ordinal
Conocimiento acerca de sistemas de diálogo	Ordinal
Experiencia usando el sistema UAH	Ordinal
Percepción de hasta qué punto UAH entiende al usuario	Ordinal
Percepción de hasta qué punto el usuario entiende a UAH	Ordinal
Velocidad de interacción percibida	Ordinal
Presencia percibida de errores cometidos por UAH	Dicotómica
Facilidad percibida de corregir los errores de UAH	Ordinal
Facilidad percibida de conseguir la información requerida	Ordinal
Satisfacción del usuario	Ordinal
Seguridad del usuario acerca de qué hacer en cada punto del diálogo	Ordinal
Percepción de comportamiento similar al humano de UAH	Ordinal
Éxito de la tarea	Dicotómica
Completitud del diálogo	Dicotómica
Duración del diálogo	Escalable
Número de turnos de usuario	Escalable
Media de palabras por turno	Escalable
WER	Escalable
Confianza de reconocimiento media	Escalable
% elocuciones correctamente comprendidas	Escalable
Número de turnos de confirmación	Escalable

Tabla 5.4. *Tipos de variables utilizadas en los estudios estadísditicos*

Todos los experimentos se realizaron utilizando el software de análisis predictivo SPSS 14[2]. Para los experimentos en los cuales el objetivo principal era obtener relaciones importantes entre todos los criterios de evaluación (tanto para parámetros de la interacción como valoraciones de calidad), se utilizaron los 37 diálogos en los cuales los usuarios contestaron al cuestionario subjetivo. En los experimentos en los cuales se estudiaron las razones que pudieron haber llevado a los usuarios a completar el cuestionario o no, se utilizaron ambos tipos de diálogos (85 en total).

[2]Statistical Product and Service Solutions - http://www.spss.com/

5.5 Resultados de la evaluación

Esta sección presenta un resumen de los resultados numéricos obtenidos a partir de los estudios estadísticos. La tabla 5.5 muestra un resumen de los resultados obtenidos con las correlaciones parciales. Por razones de espacio no se muestran las 21 tablas de correlación parcial con sus valores numéricos, sino que en su lugar, se listan únicamente las correlaciones significativas encontradas entre todas las tablas, junto con el número de tablas de correlación parcial en las cuales la relación fue significativa. Como puede observarse, no existen relaciones significativas con independencia del criterio de control empleado, es decir, ninguna de las relaciones aparece como significativa en las 21 tablas de correlación parcial.

Para cada par de criterios, la tabla 5.6 muestra el *coeficiente de correlación de Pearson* y su nivel de significación. Los niveles de significación por debajo de 0,05 se indican en azul, los que están por debajo de 0,01 se marcan en naranja, y las relaciones no significativas aparecen en blanco. De hecho, el mejor caso se alcanzó cuando la relación entre dos criterios demostró ser significativa al eliminar el efecto de 17 de las 21 variables. Este hecho demuestra que todos los criterios están estrechamente relacionados.

Relaciones entre criterios		Tablas de correlación parcial en que fue significativa
Facilidad perc. de conseguir la información requerida	Perc. de hasta qué punto UAH entiende al usuario	17
Perc. de comportamiento similar al humano de UAH	Perc. de hasta qué punto el usuario entiende a UAH	17
Conocimiento acerca de sistemas de diálogo	Conocimiento acerca de las nuevas tecnologías	17
Duración del diálogo	Número de turnos de usuario	16
Número de turnos de confirmación	Número de turnos de usuario	16
% elocuciones correctamente comprendidas	WER	16
Confianza de reconocimiento media	WER	16
Éxito de la tarea	Facilidad perc. de conseguir la información requerida	16
Facilidad perc. de conseguir la información requerida	Satisfacción del usuario	15
Perc. de comportamiento similar al humano de UAH	Satisfacción del usuario	15
Confianza de reconocimiento media	% elocuciones correctamente comprendidas	15
Éxito de la tarea	Satisfacción del usuario	15
Completitud del diálogo	Éxito de la tarea	14
Completitud del diálogo	Satisfacción del usuario	14
Facilidad perc. de corregir los errores de UAH	Facilidad perc. de conseguir la información requerida	14
Facilidad perc. de corregir los errores de UAH	Perc. de hasta qué punto UAH entiende al usuario	14
Éxito de la tarea	Perc. de hasta qué punto UAH entiende al usuario	14
Perc. de hasta qué punto UAH entiende al usuario	Satisfacción del usuario	14
WER	Avg. words per turn	14
Facilidad perc. de corregir los errores de UAH	Satisfacción del usuario	13
Facilidad perc. de conseguir la información requerida	Completitud del diálogo	13
Facilidad perc. de conseguir la información requerida	Perc. de comportamiento similar al humano de UAH	13
Completitud del diálogo	Facilidad perc. de corregir los errores de UAH	12
Completitud del diálogo	Perc. de hasta qué punto UAH entiende al usuario	12
% elocuciones correctamente comprendidas	Satisfacción del usuario	12
Facilidad perc. de corregir los errores de UAH	Éxito de la tarea	12
Perc. de comportamiento similar al humano de UAH	Éxito de la tarea	12
Perc. de comportamiento similar al humano de UAH	Perc. de hasta qué punto UAH entiende al usuario	12
Facilidad perc. de corregir los errores de UAH	Perc. de comportamiento similar al humano de UAH	11
Duración del diálogo	Facilidad perc. de conseguir la información requerida	10
Duración del diálogo	Éxito de la tarea	10
Duración del diálogo	Perc. de hasta qué punto UAH entiende al usuario	10
Duración del diálogo	Satisfacción del usuario	10
Éxito de la tarea	Número de turnos de usuario	10
Satisfacción del usuario	Perc. de hasta qué punto el usuario entiende a UAH	10
Completitud del diálogo	Duración del diálogo	9
Número de turnos de usuario	Satisfacción del usuario	7
Completitud del diálogo	Perc. de comportamiento similar al humano de UAH	6
Número de turnos de confirmación	Duración del diálogo	5
Facilidad perc. de conseguir la información requerida	Perc. de hasta qué punto el usuario entiende a UAH	5
Número de turnos de usuario	Facilidad perc. de conseguir la información requerida	4
Satisfacción del usuario	Media de palabras por turno	4
Número de turnos de confirmación	Confianza de reconocimiento media	3
Duración del diálogo	Experiencia previa con UAH	2
Completitud del diálogo	Número de turnos de usuario	1
Completitud del diálogo	Experiencia previa con UAH	1
Duración del diálogo	Facilidad perc. de corregir los errores de UAH	1
Duración del diálogo	Perc. de comportamiento similar al humano de UAH	1
Número de turnos de confirmación	Experiencia previa con UAH	1
Número de turnos de confirmación	WER	1
Número de turnos de usuario	WER	1
Facilidad perc. de conseguir la información requerida	% elocuciones correctamente comprendidas	1
Facilidad perc. de conseguir la información requerida	Confianza de reconocimiento media	1
Confianza de reconocimiento media	Éxito de la tarea	1
Confianza de reconocimiento media	Experiencia previa con UAH	1
Confianza de reconocimiento media	Seguridad del usuario acerca de qué hacer	1
Éxito de la tarea	% elocuciones correctamente comprendidas	1
Perc. de hasta qué punto UAH entiende al usuario	Número de turnos de usuario	1
Perc. de hasta qué punto UAH entiende al usuario	% elocuciones correctamente comprendidas	1
Experiencia previa con UAH	WER	1
Conocimiento acerca de sistemas de diálogo	Facilidad perc. de conseguir la información requerida	1
Conocimiento acerca de sistemas de diálogo	Perc. de hasta qué punto UAH entiende al usuario	1
Satisfacción del usuario	WER	1

Tabla 5.5. *Correlaciones parciales significativas*

1. Conocimiento NT	1																				
2. Conocimiento SDO	0,619 0,000	1																			
3. Experiencia UAH	0,116 0,506	0,226 0,191	1																		
4. UAH entiende al usuario	0,277 0,107	0,265 0,124	0,116 0,508	1																	
5. Usuario entiende a UAH	0,149 0,394	0,034 0,846	0,288 0,093	0,343 0,044	1																
6. Velocidad interacción	0,113 0,519	-0,063 0,719	0,222 0,200	0,334 0,050	0,323 0,058	1															
7. Facilidad corregir errores	0,139 0,570	-0,190 0,435	0,095 0,699	0,809 0,000	0,267 0,269	0,387 0,101	1														
8. Presencia errores	-0,034 0,848	-0,197 0,258	-0,131 0,453	-0,511 0,002	-0,258 0,135	-0,147 0,399		1													
9. Facilidad obtener info.	0,197 0,257	0,128 0,464	0,074 0,672	0,784 0,000	0,628 0,000	0,353 0,038	0,706 0,001	-0,326 0,056	1												
10. Satisfacción del usuario	0,286 0,096	0,105 0,550	0,147 0,399	0,753 0,00	0,582 0,000	0,274 0,112	0,722 0,000	-0,318 0,063	0,844 0,000	1											
11. Seguridad del usuario	0,124 0,479	0,018 0,917	0,093 0,597	0,476 0,004	0,539 0,001	0,204 0,239	0,234 0,335	-0,419 0,012	0,488 0,003	0,385 0,022	1										
12. Comportamiento humano UAH	0,238 0,168	0,208 0,230	0,284 0,098	0,684 0,000	0,764 0,000	0,443 0,008	0,601 0,006	-0,385 0,022	0,761 0,000	0,690 0,000	0,580 0,000	1									
13. Completitud del diálogo	0,113 0,518	0,003 0,987	0,212 0,222	0,485 0,003	0,321 0,060	0,152 0,382	0,623 0,004	-0,146 0,402	0,524 0,001	0,589 0,000	0,117 0,505	0,329 0,054	1								
14. Éxito de tarea	0,253 0,143	0,098 0,575	0,052 0,768	0,742 0,000	0,498 0,002	0,301 0,079	0,623 0,004	-0,226 0,191	0,912 0,000	0,827 0,000	0,331 0,052	0,637 0,000	0,614 0,000	1							
15. Duración del diálogo	-0,037 0,834	-0,072 0,682	0,104 0,554	0,433 0,009	0,155 0,375	0,215 0,215	0,421 0,073	-0,177 0,310	0,348 0,040	0,375 0,026	0,245 0,157	0,226 0,191	0,462 0,005	0,475 0,004	1						
16. Nº turnos	0,054 0,758	0,134 0,442	-0,173 0,320	0,340 0,046	-0,046 0,792	0,030 0,865	0,160 0,513	-0,193 0,266	0,249 0,149	0,257 0,137	0,183 0,293	0,027 0,878	0,354 0,037	0,328 0,054	0,744 0,000	1					
17. Confianza reconocimiento	0,152 0,384	0,149 0,394	-0,039 0,822	0,110 0,528	0,114 0,516	-0,032 0,855	0,133 0,587	-0,059 0,736	0,092 0,599	-0,024 0,891	0,315 0,065	0,246 0,154	0,070 0,688	-0,028 0,874	-0,027 0,876	0,098 0,576	1				
18, WER	-0,168 0,336	-0,034 0,845	-0,279 0,104	-0,157 0,369	-0,254 0,141	-0,134 0,443	0,081 0,742	0,191 0,273	-0,093 0,595	0,014 0,936	-0,426 0,011	-0,228 0,187	-0,163 0,349	-0,017 0,925	0,040 0,821	0,244 0,158	-0,526 0,001	1			
19. Palabras por turno	0,154 0,377	0,015 0,931	-0,057 0,745	-0,103 0,555	-0,107 0,542	-0,150 0,389	0,242 0,318	0,215 0,215	0,036 0,840	0,213 0,219	-0,158 0,364	-0,110 0,528	0,198 0,253	0,168 0,333	0,027 0,880	0,096 0,585	-0,323 0,058	0,539 0,001	1		
20. % elocuciones correct. entendidas.	0,197 0,256	0,080 0,647	0,212 0,221	0,469 0,004	0,249 0,150	-0,009 0,959	0,312 0,194	-0,341 0,045	0,350 0,040	0,495 0,002	0,398 0,018	0,379 0,025	0,291 0,089	0,329 0,054	0,115 0,510	0,027 0,878	0,440 0,008	-0,507 0,002	-0,065 0,713	1	
21. Nº turnos confirmación	0,057 0,744	0,094 0,591	-0,297 0,083	0,363 0,032	-0,027 0,878	0,000 1,000	0,227 0,350	-0,478 0,004	0,181 0,297	0,188 0,179	0,312 0,068	0,090 0,608	0,126 0,471	0,153 0,381	0,598 0,000	0,749 0,000	0,305 0,075	0,034 0,847	-0,142 0,416	0,150 0,390	1
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

Tabla 5.6. *Correlaciones entre los criterios utilizados*

Finalmente, en la tabla 5.7 se muestra un resumen de los resultados para los coeficientes *Tau-b* y *rho*, destacando las relaciones para las cuales la significación difiere de la obtenida en los estudios de correlación de Pearson. En las secciones siguientes se analizan e interpretan las conclusiones principales derivadas de estos resultados.

Criterio 1	Criterio 2	*Pearson*	*Tau-b*	*Rho*
Perc. de hasta qué punto UAH entiende al usuario	DS knowledge	0,265 0,124	0,276 0,057	0,336 0,049
Perc. de hasta qué punto UAH entiende al usuario	Perc. interaction speed	0,334 0,050	0,268 0,077	0,299 0,081
Perc. de hasta qué punto UAH entiende al usuario	Completitud del diálogo	0,485 0,003	0,390 0,013	0,426 0,011
Perc. de hasta qué punto UAH entiende al usuario	Duración del diálogo	0,433 0,009	0,209 0,111	0,278 0,105
Perc. de hasta qué punto UAH entiende al usuario	Número de turnos de usuario	0,340 0,046	0,157 0,255	0,197 0,257
Perc. de hasta qué punto UAH entiende al usuario	Número de turnos de confirmación	0,363 0,032	0,291 0,054	0,335 0,049
Perc. de hasta qué punto el usuario entiende a UAH	Éxito de la tarea	0,498 0,002	0,408 0,013	0,424 0,011
Perc. de comportamiento similar al humano de UAH	Perc. interaction speed	0,443 0,008	0,355 0,019	0,389 0,021
Perc. de comportamiento similar al humano de UAH	Facilidad perc. de corregir los errores de UAH	0,601 0,006	0,474 0,018	0,523 0,022
Completitud del diálogo	Facilidad perc. de corregir los errores de UAH	0,623 0,004	0,559 0,011	0,602 0,006
Éxito de la tarea	Facilidad perc. de corregir los errores de UAH	0,623 0,004	0,559 0,011	0,602 0,006
Facilidad perc. de conseguir la información requerida	Presencia perc. de errores cometidos por UAH	-0,326 0,056	-0,337 0,033	-0,365 0,031
Seguridad del usuario acerca de qué hacer	Presencia perc. de errores cometidos por UAH	-0,419 0,012	-0,429 0,008	-0,454 0,006
Seguridad del usuario acerca de qué hacer	Satisfacción del usuario	0,385 0,022	0,291 0,054	0,316 0,064
Duración del diálogo	Satisfacción del usuario	0,375 0,026	0,245 0,065	0,310 0,070
% elocuciones correctamente comprendidas	Satisfacción del usuario	0,495 0,002	0,223 0,151	0,248 0,151
Facilidad perc. de conseguir la información requerida	Completitud del diálogo	0,524 0,001	0,384 0,015	0,416 0,013
Duración del diálogo	Completitud del diálogo	0,462 0,005	0,350 0,014	0,421 0,012
Número de turnos de usuario	Completitud del diálogo	0,354 0,037	0,274 0,068	0,313 0,067
Facilidad perc. de conseguir la información requerida	Duración del diálogo	0,348 0,040	0,151 0,253	0,225 0,194
Duración del diálogo	Éxito de la tarea	0,475 0,004	0,362 0,011	0,435 0,009
Completitud del diálogo	Número de turnos de usuario	0,354 0,037	0,274 0,068	0,313 0,067
Seguridad del usuario acerca de qué hacer	WER	-0,426 0,011	-0,337 0,017	-0,388 0,021
Facilidad perc. de conseguir la información requerida	% elocuciones correctamente comprendidas	0,350 0,040	0,244 0,113	0,262 0,129

Tabla 5.7. *Variaciones significativas entre Pearson, Chramer's Tau-b y Spearman's Rho*

5.5.1. Influencia del desarrollo de la interacción en la decisión del usuario de responder el cuestionario subjetivo

Tal y como se ha descrito en la sección 5.3.2, no todos los usuarios contestaron el cuestionario en el cual se incluían los criterios de calidad percibida. Para estudiar si algunos de los parámetros de la interacción influenciaron el que los usuarios lo contestasen, se introdujo una variable dicotómica indicando si el usuario había rellenado o no el cuestionario y se calcularon sus correlaciones de Pearson y estudios de ANOVA con los parámetros de la interacción. La tabla 5.8 muestra los resultados obtenidos.

Relación	ANOVA F (Sig)	Eta square	Pearson (Sig)
Éxito de la tarea	7,156(0,009)	0,079	0,282(0,009)
Completitud del diálogo	7,775(0,007)	0,086	0,293(0,007)
Duración del diálogo	0,245 (0,622)	0,003	0,054 (0,622)
Número de turnos de usuario	0,729 (0,396)	0,009	0,093 (0,396)
Confianza de reconocimiento media	0,122 (0,728)	0,001	-0,159 (0,150)
WER	2,107 (0,150)	0,025	0,010 (0,927)
Media de palabras por turno	0,008 (0,927)	0,000	0,038 (0,728)
% elocuciones correctamente comprendidas	3,759 (0,056)	0,043	0,208 (0,56)
Número de turnos de confirmación	0,592 (0,447)	0,18	0,133 (0,447)

Tabla 5.8. *Significatividad entre la relación "El usuario completó el cuestionario" y los parámetros de la interacción*

Las únicas relaciones que resultaron ser significativas para la variable "el usuario completó el cuestionario" fueron con "completitud del diálogo" y "éxito de la tarea". Estos dos criterios también están correlacionados, con un valor de *ANOVA F* de 180,159, y significatividad 0,000. *Eta square* fue 0,685, y puesto que ambas son variables dicotómicas, también se calculó el valor de la *V de Cramer*, obteniendo un valor de 0,827 y significatividad 0,000.

Una conclusión que se derivará de estos resultados es que los usuarios realizaron la prueba subjetiva principalmente cuando lograron conseguir la información que requerían. El hecho de que los diálogos exitosos estuviesen correlacionados con la completitud del diálogo puede ser porque los diálogos fallidos generalmente se finalizaron prematuramente por los usuarios.

Para comprobar si los parámetros de la interacción que afectan al éxito de la tarea son los mismos para todos los grupos de usuarios, se llevaron a cabo estudios adicionales de ANOVA, con los que se obtuvieron los resultados mostrados en la tabla 5.9.

Relación	Grupo de usuarios	F	Sig
Completitud del diálogo y Éxito de la tarea	Usuarios que no completaron el test	93,312	0,000
	Usuarios que completaron el test	19,951	0,000
	Todos los usuarios	180,159	0,000
Duración del diálogo y Éxito de la tarea	Usuarios que no completaron el test	17,814	0,000
	Usuarios que completaron el test	9,638	0,004
	Todos los usuarios	21,532	0,000
Número de turnos de usuario y Éxito de la tarea	Usuarios que no completaron el test	13,025	0,001
	Usuarios que completaron el test	3,977	0,054
	Todos los usuarios	16,231	0,000
Confianza de reconocimiento media y Éxito de la tarea	Usuarios que no completaron el test	0,105	0,748
	Usuarios que completaron el test	0,026	0,874
	Todos los usuarios	0,789	0,377
WER y Éxito de la tarea	Usuarios que no completaron el test	0,171	0,681
	Usuarios que completaron el test	0,009	0,925
	Todos los usuarios	0,292	0,590
Media de palabras por turno y Éxito de la tarea	Usuarios que no completaron el test	12,787	0,001
	Usuarios que completaron el test	0,964	0,333
	Todos los usuarios	15,452	0,000
% elocuciones correctamente comprendidas y Éxito de la tarea	Usuarios que no completaron el test	5,891	0,019
	Usuarios que completaron el test	3,992	0,054
	Todos los usuarios	12,539	0,001
Número de turnos de confirmación	Usuarios que no completaron el test	0,528	0,471
	Usuarios que completaron el test	0,789	0,381
	Todos los usuarios	0,963	0,334

Tabla 5.9. *Tabla ANOVA del éxito de la tarea respecto al resto de parámetros de la interacción según los tipos de usuario*

Como puede observarse en la tabla, las únicas diferencias con respecto al éxito de la tarea aparecen para sus relaciones con el número de turnos de usuario, el porcentaje de palabras correctamente entendidas por turno, y el número de palabras por turno. Las tres relaciones fueron significativas para los usuarios que no contestaron el cuestionario, pero no para los que lo contestaron, aunque los primeros dos casos se puedan considerar prácticamente significativos en el nivel 0,05. Esta diferencia puede deberse al grado de cooperación de los diversos tipos de usuario. Por ejemplo, los usuarios que no contestaron el cuestionario y llevaron a cabo diálogos no exitosos, colgaron inmediatamente: el 70,37 % de las veces antes del cuarto turno de usuario. Sin embargo, los usuarios que contestaron a la prueba subjetiva fueron más pacientes e intentaron superar los problemas de la interacción incluso cuando al final no pudieron obtener la información que solicitaron.

La diferencia principal detectada entre los grupos de usuarios fue en la relación entre el número de palabras por turno y el éxito de la tarea. Para los usuarios que no contestaron el cuestionario, el valor de F fue 12,787, siendo significativo por debajo del nivel 0,01, mientras que para los que contestaron a la prueba, el valor de F fue 0,964 , no siendo significativo. Este hecho sucede probablemente porque la distribución del número de palabras por turno para los diálogos exitosos y no exitosos está más equilibrada en el caso de los usuarios que contestaron la prueba subjetiva. Para estos últimos usuarios, los diálogos exitosos y no exitosos tienen un número similar de palabras por turno. Sin embargo, los usuarios que no contestaron el cuestionario no emplearon más que un promedio de una palabra por turno en sus diálogos no exitosos, y más de dos turnos en los exitosos. Un promedio de palabras por turno menor o igual a una es un indicador del fallo del diálogo en el caso de los usuarios que no contestaron la prueba subjetiva.

5.5.2. Criterios con mayor influencia en la satisfacción del usuario y el éxito de la tarea

La tabla 5.10 muestra los dos valores más altos de correlación con la satisfacción del usuario, que se obtuvieron en todos los estudios estadísticos para los criterios “facilidad percibida de conseguir la información requerida” y el “éxito de la tarea”. Así, según lo esperado, la satisfacción del usuario fue alta cuando encontró facilidad para conseguir la información que requería.

Relación	*Pearson* (sig)	*Tau-b* (sig)	*Rho* (sig)	*ANOVA F* (sig)
Facilidad percibida de conseguir la información requerida y Satisfacción del usuario	0,844 (0,000)	0,750 (0,000)	0,814 (0,000)	31,071 (0,000)
Éxito de la tarea y Satisfacción del usuario	0,827 (0,000)	0,732 (0,000)	0,787 (0,000)	33,140 (0,000)

Tabla 5.10. *Significatividad estadística de las relaciones más importantes con respecto a la "satisfacción del usuario"*

Sin embargo, cabe destacar que la manera de obtener la información tiene el mismo orden de significatividad en su relación con la satisfacción del usuario como para la relación con el hecho objetivo de que finalmente se consiguiera la información. En (Möller, 2005), la satisfacción del usuario también está correlacionada con el hecho de que el usuario obtuviera finalmente la información que buscaba. Sin embargo, el indicador de Möller de la facilidad de la comunicación (que clasifica como factor de la comodidad) no proporciona una contribución significativa a la satisfacción total del usuario. Este hecho puede sugerir que la facilidad de la interacción es más importante para los usuarios que tienen una necesidad verdadera de obtener la información, en comparación con aquellos en los que la interacción se realiza siguiendo escenarios predefinidos.

Además, la pregunta del cuestionario subjetivo a partir de la cual se calcula la medida "facilidad percibida de conseguir la información requerida", considera implícitamente el éxito percibido del diálogo. Concretamente, el criterio de facilidad de la consecución de la información se obtiene de la pregunta Q8 del cuestionario (sección 5.3.2) cucyas respuestas varían de "no, fue imposible conseguir la información" a "sí, fue muy fácil conseguir la información". De este modo, existen dos criterios distintos para valorar el éxito de la tarea: un parámetro de la interacción que indica si el usuario pudo conseguir la información que buscaba ("éxito de la tarea"), y otro que indica el éxito percibido de la tarea. Este segundo criterio se extrajo del parámetro "facilidad percibida de conseguir la información requerida", asignando un 0 (fracaso) a la respuesta "no, fue imposible" y un 1 (éxito) al resto.

Las tablas de contingencia demostraron que ambas medidas del éxito de la tarea tenían el mismo valor para todos los diálogos. Por tanto, en nuestros experimentos el éxito de la tarea se considera únicamente como parámetro de la interacción. Por ejemplo, (Rajman et al., 2004) mostraron que, dado que los usuarios en evaluaciones de laboratorio no tienen la posibilidad de contrastar la información proporcionada por el sistema de diálogo, éstos confían ciegamente en las respuestas de sistema. Es decir, no comprueban si la información es correcta o útil y por tanto, consideran el hecho de obtener una respuesta del sistema equivalente a obtener un resultado correcto. Los autores estudiaron este comportamiento empleando usuarios en pruebas de laboratorio que no podían discernir si la información sobre los restaurantes, menús y precios proporcionados por un sistema de diálogo era correcta. En nuestros experimentos, se proporcionó a los usuarios de UAH información académica real. Dado que necesitaban realmente esta información, podían contrastarla y saber si era exacta o no. Así, entre los diálogos no exitosos (tanto desde el punto de vista de los parámetros de la interacción como de las valoraciones sobre la calidad) se dieron casos donde a pesar de que el sistema proporcionó al usuario información, ésta no era la que él deseaba, como demuestra el hecho de que algunos diálogos completos no fueron exitosos. La evaluación de campo presenta, de este modo, la gran ventaja de posibilitar una separación entre la calidad de la interacción y la calidad de los resultados obtenidos.

Centrándonos en los parámetros de la interacción, hay una correlación notable entre la completitud del diálogo y el éxito de la tarea. Tal y como muestra la figura 5.5, aunque los usuarios podían colgar en cuanto recibían la información deseada sin esperar a que el sistema les preguntara si deseaban alguna otra información, éstos esperaron generalmente hasta el final en los diálogos exitosos. Aunque el porcentaje de diálogos completos y exitosos fuera más alto para los usuarios más colaboradores (es decir, aquellos que contestaron el cuestionario), tanto los usuarios que llevaron a cabo la prueba subjetiva como los que no tomaron parte en ella, fueron pacientes para esperar hasta el final de los diálogos cuando eran exitosos.

Este hecho difiere de los resultados de otros autores. Por ejemplo, Turunen et al. (2006) divulgaron que había diferencias significativas entre la forma de llevar a cabo la interacción en las pruebas de laboratorio y en evaluaciones de campo con el sistema Stopman. En las pruebas de laboratorio,

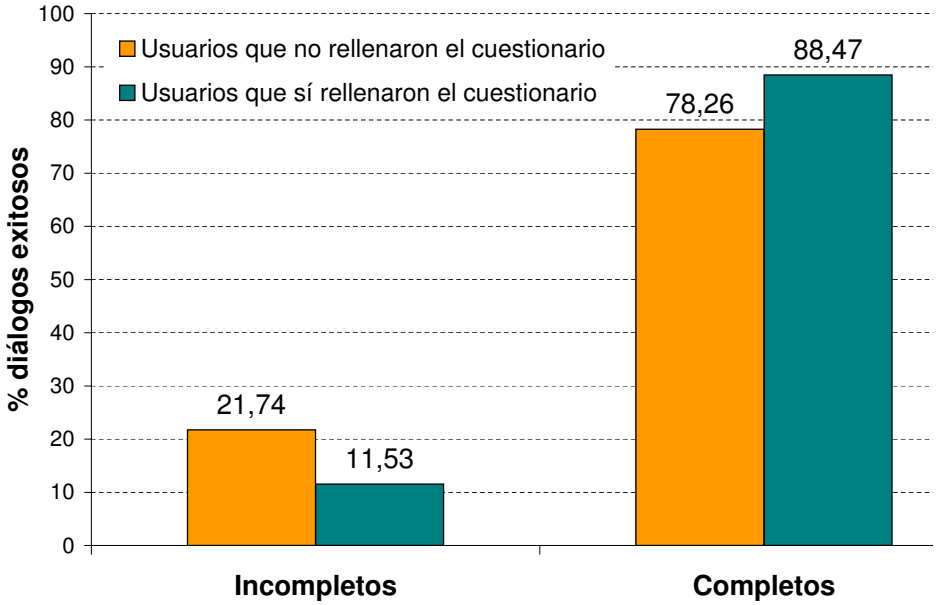


Figura 5.5. *Porcentaje de diálogos exitosos que además están completos con respecto a los diferentes grupos de usuarios*

un 65 % de los usuarios emplearon una petición explícita de terminación de llamada (por ejemplo, "gracias y adiós"). Por el contrario, en la evaluación de campo al menos de un 10 % de los usuarios esperaron al final de la llamada antes de colgar. El número de diálogos en los cuales los usuarios esperaron hasta el final de la interacción (es decir el número de diálogos completos) en nuestro estudio de campo es un 50 % mayor que el mostrado en (Turunen et al., 2006).

Rajman et al. (2004) argumentan que el hecho de que los usuarios demuestren una actitud positiva hacia el sistema no sólo depende del funcionamiento de éste sino también del perfil "tecnofóbico" o "tecnofílico" de los usuarios, aunque no controlaron este parámetro en su experimentación. En nuestros experimentos, un 57 % de los usuarios clasificaron su conocimiento sobre las nuevas tecnologías para el acceso a la información con 3 sobre una escala entre 1 y 5, donde 1 representaba "bajo" y 5 "alto". De este modo, la colaboración de nuestros usuarios podría ser como resultado de su disposición favorable hacia las nuevas tecnologías.

Otro criterio que está estrechamente correlacionado con el éxito de la tarea y la satisfacción del usuario es la facilidad percibida para corregir errores. Sin embargo, la presencia percibida de errores no se correla con nin-

guno de estos criterios. Esto puede deberse a que, aunque en el 48,19 % de los diálogos exitosos los usuarios detectaron errores, en la mayoría de los casos supieron corregirlos y obtener la información que buscaban. Concretamente, según indica la figura 5.6, un 69,23 % de los usuarios encontraron "fácil" o "muy fácil" corregir errores en los diálogos exitosos. Sin embargo, en los no exitosos, un 83,33 % de los usuarios manifestó que la corrección de errores fue "difícil" o "muy difícil".

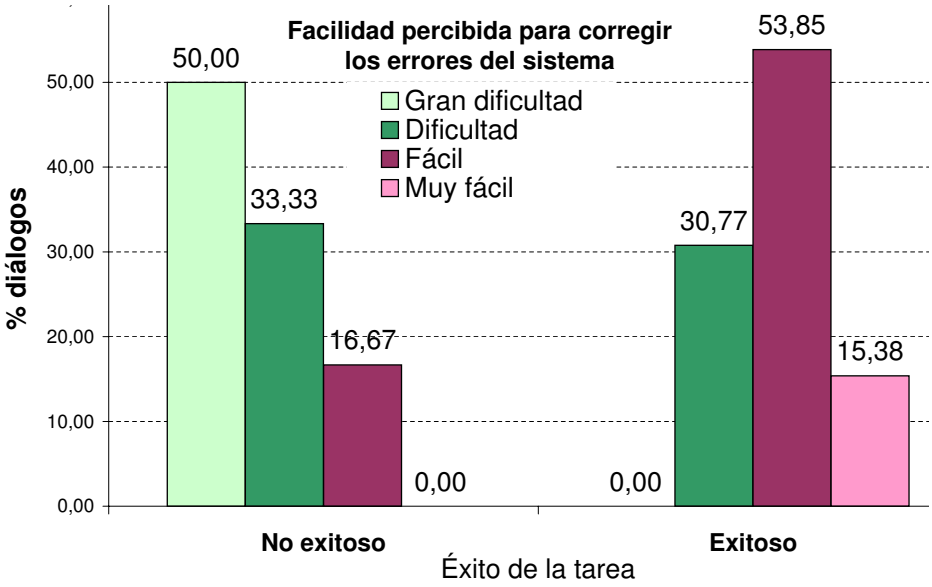


Figura 5.6. *Éxito de la tarea con respecto a la percepción de la facilidad de corregir errores*

En (Möller, 2005), la opinión de los usuarios sobre si los malentendidos podrían ser aclarados fácilmente (que se clasificó como un factor que contribuía a la calidad del diálogo), no resultó ser un buen indicador de la satisfacción del usuario. Además, el autor encontró que la satisfacción del usuario no se podía predecir completamente mediante el éxito de la tarea, y sostuvo que este resultado podría ser debido a las condiciones poco realistas de la experimentación de laboratorio empleada en su investigación. En la tesis doctoral se ha corroborado este hecho en un estudio de campo, puesto que los cuestionarios subjetivos no se pudieron substituir por los parámetros de la interacción empleados sin que esto supusiera pérdida de información.

5.5.3. Criterios que poseen el mayor número de relaciones significativas

El criterio que mostró un mayor número de correlaciones significativas fue la "percepción de hasta qué punto UAH entiende al usuario". Por una parte, está altamente correlacionado con otras valoraciones de calidad, como el grado hasta el cual el usuario entiende al sistema, la facilidad percibida de la corrección de errores, la facilidad percibida de obtener la información, la satisfacción del usuario, la presencia percibida de errores (correlación negativa en este caso), y la percepción de un comportamiento parecido al humano por parte del sistema. Además, como puede observarse en la tabla 5.6, en la mayor parte de estas relaciones la significatividad fue máxima. Por otra parte, este criterio de calidad está altamente correlacionado con parámetros de la interacción como el haber completado el diálogo, el éxito de la tarea, la duración del diálogo o el porcentaje de elocuciones correctamente entendidas por diálogo.

Las relaciones más significativas entre esta valoración de la calidad y otros parámetros se obtuvieron con el éxito de la tarea y la satisfacción del usuario (figura 5.7). Un ajuste linear mostró un coeficiente de determinación múltiple de 0,55, lo que indica que el 55 % de la variabilidad de la comprensión percibida por parte de UAH se podría explicar mediante el éxito de la tarea. La comprensión percibida del sistema, que es mencionada en Möller (2005) como indicador de la calidad de la entradas al sistema, también está correlacionada de manera perceptible con la satisfacción del usuario en el estudio de Möller.

Cabe también destacar que el grado con el cual el usuario percibió que el sistema UAH le entendía no está correlacionado con los parámetros de la interacción que miden el funcionamiento del reconocedor del habla, como el WER o las medidas de confianza. Sin embargo, si que está correlacionado con el porcentaje de elocuciones correctamente entendidas con una significatividad de 0,01; lo que indica que desde el punto de vista del usuario, los errores de reconocimiento del habla no fueron importantes mientras las interpretaciones semánticas fuesen correctas y estos errores fuesen imperceptibles por el usuario. Este hecho se refleja en que la presencia percibida de errores está relacionada con el porcentaje de elocuciones correctamente entendidas y el número de turnos de confirmación, pero no con el WER. Sin embargo, la facilidad percibida de corregir errores no está correlacionada con ninguna

de estas medidas. La percepción de presencia de errores y la facilidad percibida de corregirlos están significativamente correlacionadas con la opinión del usuario acerca de cómo le entendía el sistema. La presencia percibida de errores también afecta negativamente a la seguridad del usuario acerca de qué debe decir o cómo debe comportarse durante la interacción.

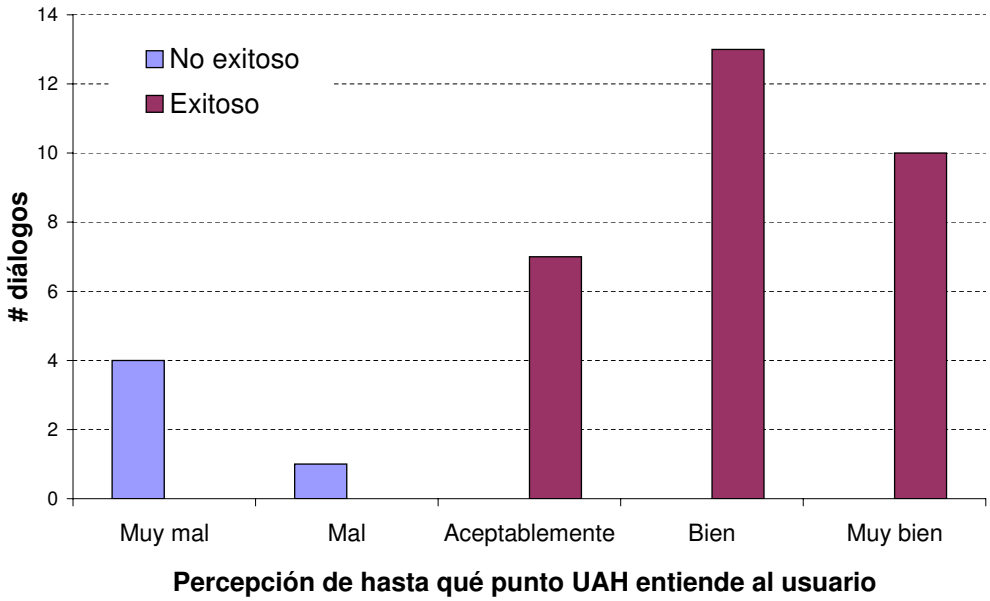


Figura 5.7. *Relación entre la percepción de hasta qué punto UAH entiende al usuario y el éxito de la tarea*

5.5.4. Influencia del conocimiento previo y la experiencia del usuario

Es significativo que el conocimiento que poseían los usuarios sobre los sistemas de diálogo y las nuevas tecnologías para el acceso a la información tuviesen los factores de correlación más bajos con el resto de criterios, salvo entre sí. De este modo, en nuestro caso, el conocimiento de los usuarios sobre las nuevas tecnologías para el acceso a la información no fue determinante en los resultados de la interacción, ni en términos objetivos (p.ej. duración y éxito), ni en las valoraciones percibidas (p.ej. velocidad y satisfacción percibidas por el usuario). Esta situación puede ser una consecuencia del alto

nivel de conocimiento técnico que la mayoría de los usuarios de UAH dijeron tener. Es posible que en experimentos con otros sistemas de diálogo donde los usuarios tengan niveles de conocimiento más variados, éste pueda ser un criterio importante.

La experiencia previa del usuario utilizando el sistema ("empleo previo de UAH") tampoco está correlacionada significativamente con los demás criterios en ninguno de los estudios estadísticos llevados a cabo. Sin embargo, el signo de los parámetros de la correlación indica que los usuarios experimentados percibieron menos errores, necesitaron menos turnos para conseguir la información, originaron menos errores de reconocimiento y requirieron menos turnos de confirmación que los usuarios noveles.

El hecho de que el uso previo de UAH no estuviera correlacionado con otros factores, como el éxito de la tarea o la velocidad de la interacción, difiere de los resultados encontrados en la literatura. Turunen et al. (2006) indicaron que la experiencia anterior al utilizar un sistema es un factor de relevancia que puede ayudar a predecir el éxito y la calidad del diálogo. Park et al. (2007) demostró que el funcionamiento del sistema es mejor para usuarios que lo han empleado con anterioridad. Otros autores han estudiado el efecto de la experiencia del usuario sobre la valoración de la calidad. Sturm et al. (2005) indican que un uso previo del sistema de manera prolongada permite obtener mejoras substanciales en las valoraciones de calidad, tales como la facilidad de uso y la satisfacción del usuario.

Creemos que el impacto de la experiencia del usuario está estrechamente vinculado al tipo de evaluación realizada. En las pruebas de laboratorio, se suele entrenar a los usuarios sobre cómo utilizar el sistema, o al menos, se les suele informar acerca de cómo interactuar con él. En los estudios de campo, comúnmente los usuarios emplean el sistema sin ningún entrenamiento anterior, y por esta razón son menos propensos a emplear funcionalidades como la petición de ayuda (Turunen et al., 2006), de la que muchas veces no son conscientes, aunque estas características sean muy útiles para facilitar la interacción y poder recuperarse de situaciones de error. Por otra parte, en algunos campos de estudio, como la utilización de los sistemas de diálogo oral en el ámbito sanitario, se ha discutido que, contrariamente a lo que sugieren los estudios previamente comentados, la experiencia previa utilizando el sistema no implica siempre un mejor funcionamiento y una percepción más positiva de la calidad del sistema. Bickmore y Giorgino (2006) muestran

que los usuarios que utilizan intermitentemente sistemas de diálogo sanitario telefónica, comparados con aquellos que los utilizan con frecuencia y los que no los utilizan apenas, obtienen niveles de satisfacción más altos y mejores resultados en cuanto a ventajas percibidas. Sin embargo, como muestra Farzanfar et al. (2004), este hecho puede ser debido a la tensión que algunos usuarios experimentan al sentirse supervisados.

5.5.5. Influencia de la iniciativa de gestión del diálogo

Para estudiar la influencia de la iniciativa utilizada para la gestión del diálogo, se repitió de nuevo la experimentación detallada anteriormente, pero distinguiendo entre los diálogos con iniciativa dirigida por el sistema y los diálogos con iniciativa mixta. Las diferencias entre ambas aproximaciones se muestran en la tabla 5.11, donde las correlaciones significativas aparecen indicadas con una 'S' (sí) y las no significativas con una 'N' (no).

Se observa que el éxito de la tarea es aproximadamente igual para ambas iniciativas de gestión del diálogo. Este resultado difiere de los que se pueden encontrar en la literatura [3], donde una iniciativa más flexible conduce a tasas de éxito considerablemente más altas. En nuestros experimentos el éxito fue mayor para la iniciativa mixta, pero la diferencia entre ambas es prácticamente insignificante (el 77,77 % de los diálogos de iniciativa mixta y el 76,92 % de los diálogos con iniciativa por parte del sistema concluyeron con éxito).

[3]Möller (2005) presenta un resumen muy detallado.

Criterio 1	Criterio 2	Inic. mixta	Inic. sistema
Perc. de hasta qué punto el usuario entiende a UAH	Perc. de hasta qué punto UAH entiende al usuario	N	S
Perc. interaction speed	Perc. de hasta qué punto UAH entiende al usuario	S	N
Presencia perc. de errores cometidos por UAH	Conocimiento acerca de sistemas de diálogo	S	N
Presencia perc. de errores cometidos por UAH	Perc. de hasta qué punto UAH entiende al usuario	N	S
Seguridad del usuario acerca de qué hacer	Presencia perc. de errores cometidos por UAH	N	S
Seguridad del usuario acerca de qué hacer	Facilidad perc. de conseguir la información requerida	N	S
Seguridad del usuario acerca de qué hacer	Satisfacción del usuario	N	S
Perc. de comportamiento similar al humano de UAH	Presencia perc. de errores cometidos por UAH	N	S
Perc. de comportamiento similar al humano de UAH	Facilidad perc. de conseguir la información requerida	N	S
Perc. de comportamiento similar al humano de UAH	Satisfacción del usuario	N	S
Perc. de comportamiento similar al humano de UAH	Seguridad del usuario acerca de qué hacer	N	S
WER	Presencia perc. de errores cometidos por UAH	N	S
Éxito de la tarea	Seguridad del usuario acerca de qué hacer	N	S
Éxito de la tarea	Perc. de comportamiento similar al humano de UAH	N	S
Éxito de la tarea	Completitud del diálogo	S	N
Duración del diálogo	Completitud del diálogo	N	S
Duración del diálogo	Facilidad perc. de conseguir la información requerida	S	N
Duración del diálogo	Satisfacción del usuario	S	N
Duración del diálogo	Éxito de la tarea	S	N
Número de turnos de usuario	Éxito de la tarea	S	N
Número de turnos de usuario	Satisfacción del usuario	S	N
Número de turnos de usuario	Facilidad perc. de conseguir la información requerida	S	N
Completitud del diálogo	Facilidad perc. de conseguir la información requerida	N	S
Confianza de reconocimiento media	Satisfacción del usuario	S	N
WER	Seguridad del usuario acerca de qué hacer	N	S
WER	Completitud del diálogo	N	S
WER	Número de turnos de usuario	S	N
% elocuciones correctamente comprendidas	Facilidad perc. de conseguir la información requerida	N	S
% elocuciones correctamente comprendidas	Satisfacción del usuario	N	S
% elocuciones correctamente comprendidas	Seguridad del usuario acerca de qué hacer	N	S
% elocuciones correctamente comprendidas	Perc. de comportamiento similar al humano de UAH	N	S
% elocuciones correctamente comprendidas	Completitud del diálogo	N	S
% elocuciones correctamente comprendidas	Éxito de la tarea	N	S
% elocuciones correctamente comprendidas	Confianza de reconocimiento media	N	S
Número de turnos de confirmación	Presencia perc. de errores cometidos por UAH	N	S
Número de turnos de confirmación	Duración del diálogo	N	S
Número de turnos de confirmación	Número de turnos de usuario	N	S
Número de turnos de confirmación	Confianza de reconocimiento media	N	S

Tabla 5.11. *Criterios que aparecen correlacionados significativamente con una iniciativa del diálogo pero no con la otra*

Sin embargo, a la luz de los resultados experimentales, el éxito de la tarea parece estar relacionado con distintos factores en cada tipo de iniciativa. De esta manera, en la iniciativa mixta la seguridad del usuario sobre qué hacer en cada momento del diálogo no está correlacionado con el éxito de la tarea, la satisfacción del usuario ni la percepción sobre la facilidad de obtener la información requerida. Por el contrario, el éxito de la tarea tiene una correlación significativa con la seguridad del usuario en los diálogos dirigidos por el sistema. Este hecho sucede probablemente porque el usuario dispone de mayor libertad en las interacciones con iniciativa mixta y, por tanto, no se restringe lo que debe decir en cada momento (figura 5.8). Sin embargo, esta situación no condujo a malos resultados de la interacción, pues el éxito de la tarea no se redujo al emplear iniciativa mixta.

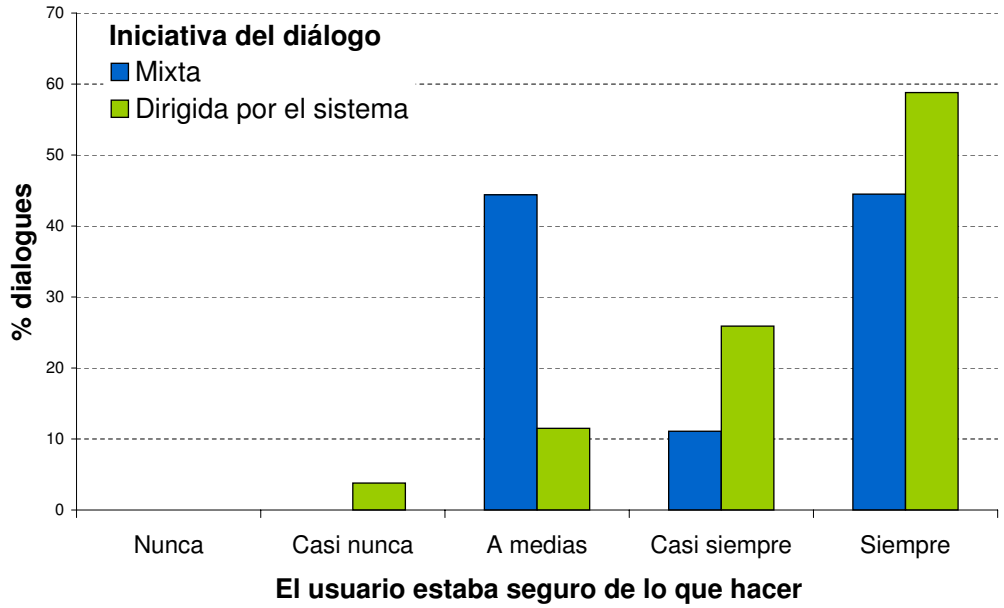


Figura 5.8. *Influencia de la iniciativa del diálogo en la seguridad del usuario acerca de qué hacer*

Las correlaciones de la facilidad percibida de conseguir la información requerida fueron también muy diferentes en ambos casos. En el caso de la iniciativa por parte del sistema está relacionada con la completitud del diálogo, el porcentaje de elocuciones correctamente comprendidas y la opinión que el usuario tenía sobre el comportamiento humano del sistema. Por el contrario, para los diálogos con iniciativa mixta la facilidad percibida no está

correlacionada con estas medidas, sino con indicadores de la duración de la interacción tales como la "duración del diálogo" o el "número de turnos de usuario". Igualmente sucede con la satisfacción y el éxito de la tarea, que están altamente correlacionadas con medidas de duración en interacciones con iniciativa mixta, pero no en los diálogos dirigidos por el sistema. La duración de estos diálogos está correlacionada de manera perceptible con la satisfacción del usuario, mientras que en sistemas con iniciativas más estrictas de la interacción no fue considerada tan importante por los usuarios. Además, como puede observarse en la figura 5.9, la duración media de los diálogos es menor cuando la interacción es más flexible (iniciativa mixta en lugar de dirigida por el sistema).

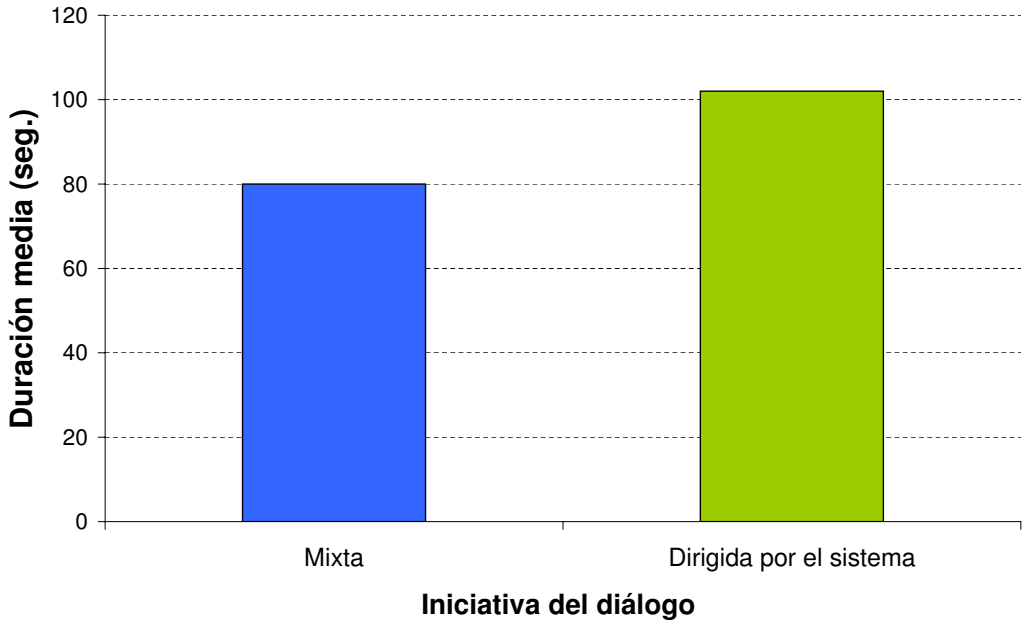


Figura 5.9. *Duración del diálogo para cada una de las iniciativas de gestión del diálogo*

Adicionalmente, la percepción de errores está relacionada en los diálogos con iniciativa mixta con el conocimiento del usuario acerca de los sistemas de diálogo. Este no fue el caso de la iniciativa dirigida por el sistema. Además, no está correlacionado con otras medidas como la confianza del usuario, el WER o el número de turnos de confirmación, que mostraron ser factores importantes en los diálogos dirigidos por el sistema.

Los estudios basados en pruebas de laboratorio como los de Rajman et al. (2004) no han podido percibir variaciones claras en la calidad con respecto al predominio de la iniciativa del sistema o del usuario. Además, algunas pruebas de laboratorio como las llevadas a cabo en (Möller, 2005) para el sistema BoRIS no pudieron encontrar ninguna relación significativa entre la iniciativa y otros parámetros de la interacción. Sin embargo, nuestros resultados demuestran que la significación de las relaciones entre los diversos criterios de evaluación, incluyendo parámetros de la interacción y valoraciones de la calidad, varía dependiendo de la iniciativa utilizada para la gestión del diálogo.

5.6 Conclusiones

En este capítulo se ha presentado un estudio de las relaciones entre varios criterios estándar de-facto para la evaluación de un sistema de diálogo oral que utiliza el canal telefónico. Nuestros resultados experimentales se basan en un estudio de campo que utiliza interacciones reales registradas por usuarios no reclutados previamente que llamaron espontáneamente al sistema para obtener información.

Para realizar nuestro estudio se han calculado parámetros de la interacción (o medidas objetivas) y juicios de la calidad (medidas subjetivas) empleando un corpus de las interacciones reales sistema-usuario. En concreto, los criterios cuantitativos empleados han sido: duración del diálogo, completitud del diálogo, éxito de la tarea, número de turnos de usuario, media de palabras por turno, confianza de reconocimiento media, WER, porcentaje de elocuciones correctamente comprendidas y número de turnos de confirmación. Las medidas cualitativas se extrajeron de cuestionarios que los usuarios podían completar opcionalmente. Los criterios empleados han sido: el grado en el cual el usuario valoraba que era entendido correctamente por el sistema, el grado en el cual el usuario entendía los mensajes del sistema, la velocidad percibida de la interacción, la facilidad percibida de la corrección de errores, la presencia percibida de errores, la seguridad del usuario acerca de lo que debía hacer en cada momento de la interacción, el grado en el cual el usuario creyó que el comportamiento de sistema era similar al humano y el nivel de satisfacción del usuario con la interacción. También se consideró información adicional de los usuarios: su conocimiento sobre las nuevas tecnologías para

el acceso a la información, su conocimiento sobre los sistemas de diálogo oral y el número de veces que el usuario había utilizado previamente el sistema.

Se han llevado a cabo un conjunto de estudios estadísticos a partir de los cuales se han extraído relaciones significativas entre todos los criterios. Esta aproximación no se ha desarrollado suficientemente en la literatura, y en la tesis doctoral se han obtenido algunos resultados empíricos notables. Nuestra evidencia demuestra que el éxito de la tarea, la facilidad percibida de obtener la información y el punto hasta el cual el usuario percibe que el sistema lo entiende están correlacionados con la satisfacción del usuario. Estos resultados sugieren que obtener la información requerida no conlleva necesariamente la satisfacción del usuario, dado que los usuarios valoraron en algunos casos que diálogos exitosos no les habían satisfecho debido a que encontraron dificultades para obtener la información que estaban buscando (a pesar de haber recibido datos que concordaban con su petición). Ésta es una de las implicaciones derivadas del uso de los estudios de campo, en los cuales los usuarios se preocupan no sólo de obtener la información que buscaban, sino también de obtenerla fácilmente y de que ésta sea correcta. Además, la relación entre la facilidad percibida de obtener la información y otros criterios varía notablemente con la estrategia de gestión de diálogo empleada. Nuestros resultados experimentales demuestran que, en los diálogos dirigidos por el sistema, la facilidad percibida está relacionada con el buen funcionamiento del módulo de comprensión. Por el contrario, en diálogos con iniciativa mixta, la satisfacción del usuario y la facilidad percibida de obtener la información parecen estar relacionadas principalmente con métricas de duración. Este hecho tiene una fuerte implicación en las valoraciones sobre la calidad, pues el éxito de la tarea está altamente correlacionado con la satisfacción del usuario en ambas iniciativas. Así, nuestros resultados sugieren que la predicción de la satisfacción del usuario también depende de la iniciativa del diálogo empleada. En los diálogos con iniciativa mixta parece estar relacionada más directamente con medidas objetivas, como la duración del diálogo. Sin embargo, en diálogos más restringidos, las medidas subjetivas como el grado hasta el cual el usuario percibe que el sistema le entiende, tienen un impacto mayor. Se trata de un resultado importante que podría indicar una necesidad de adaptar los procedimientos de evaluación al tipo de interacciones que se analizan.

Además, se han estudiado los motivos que hicieron a algunos usuarios contestar a la prueba subjetiva opcional a partir de la cual se extrajeron las valoraciones sobre la calidad. Se ha determinado que puede explicarse principalmente en términos de la completitud del diálogo y del éxito de la tarea. De este modo, los experimentos realizados incluyendo las opiniones de los usuarios sobre la calidad del sistema, se correspondieron principalmente a los diálogos finalizados con éxito, en los cuales los usuarios obtuvieron la información que buscaban. Éste es uno los motivos por los que se determinó que estos usuarios eran muy cooperativos, lo que explica el alto porcentaje de diálogos completados. Además, contrariamente a lo sucede generalmente en los estudios de laboratorio, estas medidas consideran que incluso cuando el usuario obtiene la información del sistema, el diálogo no puede considerarse exitoso si la información proporcionada no es correcta. Finalmente, no se encontró ninguna evidencia del efecto de la experiencia anterior de los usuarios empleando el sistema sobre el rendimiento del mismo o sobre el éxito de la tarea.

Hemos llegado al fin y yo inauguro
triste mi paz: la obra está completa.
Jorge Guillén, Obra completa

6 Conclusiones y trabajo futuro

6.1 Resumen de las contribuciones

La mayoría de los expertos coinciden en que los sistemas de diálogo futuros deben ser portables entre idiomas, dominios y plataformas, así como tener la capacidad de adaptarse a los usuarios y al contexto de la comunicación. En la presente tesis se han realizado diversas contribuciones para posibilitar la adaptación de los sistemas de diálogo al contexto de la interacción y el estado emocional de los usuarios, así como su adaptación a las expectativas y necesidades específicas de los mismos extraídas mediante estudios de evaluación de campo. Por último, se facilita la portabilidad de reconocedores del habla al funcionamiento en distintos idiomas, sin la necesidad de generar nuevos recursos para cada uno de ellos.

Para evaluar las contribuciones de la tesis se ha desarrollado el **sistema de diálogo UAH** (capítulo 2), que proporciona información académica sobre el Departamento de Lenguajes y Sistemas Informáticos. El sistema sigue una arquitectura compuesta por cinco módulos: un reconocedor del habla, un módulo de generación automática de gramáticas, un gestor de diálogo, un módulo de acceso a la base de datos y un generador oral de respuestas. Entre ellos cabe destacar **el módulo de generación automática de gramáticas (GAG)**. Puesto que la información que aporta UAH se modifica continuamente en la base de datos donde está almacenada, el módulo GAG sigue un método para actualizar las gramáticas de reconocimiento con la última versión contenida en las bases de datos sin introducir retardos en la interacción. Esta técnica se ha denominado **"creación de reglas gramaticales mediante disparadores" (TGC)**. El sistema incluye además

diversas técnicas de confirmación (explícitas e implícitas), así como **varias iniciativas de gestión del diálogo** (dirigida por el sistema y mixta), con la finalidad de estudiar las ventajas de cada una de ellas en cuanto al funcionamiento real del sistema y su usabilidad percibida. La gestión del diálogo se llevó a cabo en UAH mediante documentos VoiceXML **generados dinámicamente dependiendo del contexto de la interacción**, información que se empleó asimismo para adaptar la salida oral del sistema a las necesidades del usuario.

El sistema UAH se puso a disposición pública en junio de 2005. **A partir de las interacciones de los usuarios con el sistema se ha generado y etiquetado de forma semi-automática un corpus de evaluación**. Las elocuciones de los usuarios se almacenaron en formato WAV junto con información acerca de la hora de comienzo y fin de la grabación, la respuesta anterior del sistema y el resultado del reconocimiento del habla (incluyendo las medidas de confianza). Esta información se almacenó en una base de datos conjuntamente con la hora de inicio y final del diálogo. A partir de esta información, se calcularon nueve criterios que posteriormente se emplearían para evaluar el sistema, algunos se obtuvieron automáticamente (p.ej. la duración del diálogo) y otros requirieron anotación humana (p.ej. éxito del diálogo). Siguiendo esta metodología se generó **un corpus etiquetado de 85 diálogos (422 turnos de usuario)** correspondiente a las llamadas recibidas en el sistema durante un año. Adicionalmente, se invitó a los usuarios a completar un cuestionario en el cual podrían dar su opinión personal sobre diversos aspectos de la interacción con el sistema. A partir de las respuestas al cuestionario se obtuvieron un total de 12 criterios de calidad para cada elocución, incluyendo parámetros como la satisfacción del usuario o la velocidad de interacción percibida. Por último, se realizó un etiquetado adicional del corpus para clasificar cada una de las elocuciones en una de las siguientes categorías emocionales: *neutro*, *duda*, *enfado* y *aburrimiento*. En este proceso participaron nueve anotadores no expertos que etiquetaron el corpus dos veces (con y sin información contextual).

En el capítulo 3 se han descrito las aportaciones de la tesis con respecto al reconocimiento de emociones. En primer lugar, se ha presentado un **estado de el arte** que describe las principales aproximaciones disponibles en la literatura y que revela cómo la investigación en este área se centra principalmente en la forma de aplicar diversos algoritmos de aprendizaje au-

tomático para distinguir entre emociones, prestándose una menor atención a determinar en qué información debe basarse dicho proceso. El enfoque tradicional se basa en emplear información multimodal (acústica y visual) obtenida de emociones actuadas, con el fin de tener varias modalidades en las que basar el reconocimiento así como un control estricto sobre los datos recogidos. La investigación realizada en la tesis tiene como una de las aportaciones principales, **la inclusión de información contextual para el reconocimiento de emociones no actuadas en sistemas de diálogo oral**. Éste ha sido un reto ambicioso debido, en primer lugar, a que el reconocimiento debía basarse en una única modalidad de entrada y, en segundo lugar, a que las emociones naturales se expresan muy sutilmente y en ellas las categorías emocionales definidas están generalmente muy desequilibradas (hay una emoción predominante, que generalmente se corresponde con el estado "neutro"), lo que plantea diversas dificultades que, a nuestro entender, no habían sido tratadas por completo en la literatura.

Una de las principales dificultades era la obtención de anotaciones fiables de los corpus emocionales a pesar de que las medidas de acuerdo entre anotadores se ven afectadas profundamente por el sesgo de los corpus no actuados, otra era que debido al desequilibrio del corpus, es difícil encontrar métodos automáticos de clasificación que superen a un baseline que clasifique siempre con la categoría que se da mayoritariamente. En el capítulo 3 se ha presentado **una discusión detallada sobre cómo calcular e interpretar con fiabilidad los coeficientes kappa** en corpus de emociones reales. La metodología propuesta se ha **evaluado experimentalmente utilizando el corpus emocional UAH, midiéndose la significatividad estadística de los resultados**. Los resultados obtenidos muestran que, por una parte, **la aproximación propuesta permite obtener valores de acuerdo entre anotadores más cercanos al máximo alcanzable, posibilitando que anotadores no expertos puedan detectar un mayor número de emociones no neutras, y que el resultado de la anotación se vea afectado en menor medida por las diferencias entre anotadores**. Por otra parte, evidencian un **40 % de mejora en el reconocimiento automático de emociones respecto a los resultados obtenidos con las aproximaciones de la literatura** (basadas únicamente en las características acústicas sin considerar información contextual). La evaluación de la metodología mediante corpus de emociones reales **posibilita su aplicación**

práctica, dado que todos los métodos propuestos se pueden utilizar durante la ejecución de los sistemas de diálogo.

En el capítulo 4 se ha presentado el trabajo realizado durante una estancia investigadora de tres meses en el Laboratory of Computer Speech Processing de la Technical University of Liberec bajo la supervisión del profesor Jan Nouza. Tal y como se describe en el capítulo, el aprovechamiento de los recursos disponibles en un idioma para llevar a cabo el reconocimiento de un idioma distinto facilita el desarrollo de reconocedores del habla, especialmente cuando se trata de idiomas o dialectos para los que se dispone de un número limitado de recursos. Se ha desarrollado un método eficiente para la adaptación de recursos disponibles en un idioma para el reconocimiento de otros idiomas distintos. Las aproximaciones que se pueden encontrar en la literatura requieren generalmente estudios fonéticos y lingüísticos muy complejos de los idiomas implicados, mientras que el método propuesto en la tesis permite realizar la adaptación de forma rápida y eficiente. La aproximación propuesta se ha evaluado experimentalmente utilizando el sistema MyVoice, desarrollado por la Technical University of Liberec para usuarios discapacitados checos, y cuya **adaptación al español** se presenta en la tesis. Se ha demostrado empíricamente que la adaptación entre idiomas propuesta se puede llevar a cabo en un breve periodo de tiempo mediante correspondencias diseñadas por expertos entre los alfabetos fonéticos de ambos idiomas. Los resultados experimentales muestran que, para una tarea con un vocabulario de 432 palabras, puede lograrse una precisión del 95,6 % para el español y del 97,5 % para el eslovaco después de adaptar un reconocedor de checo. Para vocabularios de hasta 149k palabras, se obtienen resultados del 72,9 % para el español y 77,4 % para el eslovaco. Se trata de resultados muy prometedores, puesto que demuestran que **la portabilidad de los reconocedores del habla se puede realizar de una forma directa por medio de nuestra aproximación, lográndose buenos resultados tanto con idiomas muy similares (como el checo y el eslovaco), como para idiomas fonéticamente muy distintos (como el checo - lengua eslava - y el español - lengua itálica)**.

En el capítulo 5 se ha descrito **la evaluación de campo del sistema UAH**. En la literatura la evaluación de sistemas de diálogo se realiza generalmente bajo condiciones estrictas de laboratorio en las cuales se reclutan usuarios para interactuar con los sistemas siguiendo una lista predefinida

de escenarios. La desventaja principal de este método radica en que los escenarios pueden diferenciarse de las tareas que un usuario seleccionaría en una interacción real con el sistema. Para la tesis se ha realizado un estudio de campo en el que los usuarios interactuaron con el sistema por iniciativa propia y en el que, por tanto, se contempla la necesidad real del usuario de obtener la información que el sistema proporciona. Por otra parte, la mayoría de las aproximaciones de evaluación en la literatura contemplan por separado los parámetros de la interacción y la valoración subjetiva de la calidad; mientras que nuestros resultados miden la conexión entre ambos determinando las relaciones significativas. Para ello se emplean diversos coeficientes estadísticos adaptados al tipo de información procesada, como por ejemplo *coeficientes de correlación de Pearson*, *estudios de ANOVA*, *tau-b*, *rho* y *coeficientes Eta-cuadrado*. Nuestro estudio proporciona **nuevas evidencias empíricas** que son relevantes para predecir comportamientos del sistema y que han arrojado luz acerca de los criterios que afectan en mayor medida a la satisfacción del usuario y al éxito de la tarea, estudiando además el impacto del grado de conocimiento previo y experiencia del usuario y la utilización de diversas iniciativas de gestión del diálogo. La utilización de estas relaciones empíricas mejora el desarrollo y la evaluación de los sistemas y la predicción de su rendimiento o usabilidad.

6.2 Trabajo futuro

6.2.1. Reconocimiento de emociones no actuadas

Los planes de trabajo futuro incluyen evaluar la técnica presentada utilizando otros corpus, para garantizar así su independencia del dominio de aplicación. Adicionalmente, se ha comenzado ya con la integración del reconocedor de emociones en el sistema UAH, se prevé que éste funcione en paralelo con el reconocedor del habla, aceptando como entrada la señal de voz obtenida con la tarjeta telefónica, y generando como salida la emoción reconocida, que será transmitida al gestor de diálogo para que adapte su comportamiento a la misma. El objetivo a priori es utilizar solamente las emociones descritas en la tesis: enfado, aburrimiento y duda. La gestión del diálogo para los usuarios dubitativos requerirá proporcionarles ayuda detallada y usar iniciativas dirigidas por el sistema que permitan clarificar al

usuario qué información debe suministrar en ese paso de la interacción. Por el contrario, si se detecta que el usuario está aburrido, la estrategia óptima podría ser acortar las intervenciones del sistema, cambiar la prosodia de la síntesis de voz y emplear confirmaciones implícitas. Para el estado emocional *enfadado*, deben considerase estrategias de recuperación de errores y técnicas que permitan transmitir al usuario el origen de los malentendidos que puedan producirse.

6.2.2. Adaptación de reconocedores del habla entre idiomas

Nuestra línea de trabajo futuro más inmediata consiste en llevar a cabo una comparación entre el reconocedor checo adaptado al eslovaco y al español y otros sistemas de reconocimiento de habla ideados para estos dos últimos idiomas. Ya hemos realizado experimentos preliminares con el reconocedor del habla en español de Windows Vista alcanzando una tasa de error que es tan sólo un 4 % absoluto menor que la obtenida para el sistema adaptado. Este hecho indica que los resultados del método propuesto no difieren excesivamente de los que se podrían alcanzar para un sistema desarrollado originalmente en la lengua objetivo. Para obtener comparativas más fiables desarrollaremos experimentos adicionales en esta línea de trabajo.

Por otra parte, los resultados prometedores mostrados en el capítulo 4 nos animan a considerar el uso futuro de nuestra adaptación fonética para el caso de idiomas con recursos lingüísticos muy limitados. Una tarea particularmente interesante sería estudiar la conveniencia de la técnica propuesta para la adaptación a idiomas minoritarios que no pertenecen a la familia indo-europea.

6.2.3. Evaluación de campo de sistemas de diálogo oral

Además de los diálogos empleados en el capítulo 5, se ha adquirido otro corpus bajo condiciones de laboratorio, concretamente mediante usuarios reclutados que utilizaron escenarios predefinidos para interactuar con el sistema UAH. Nuestro objetivo es llevar a cabo una evaluación de este corpus

y comparar los resultados de la misma con aquellos obtenidos mediante la evaluación de campo descrita en la tesis.

Creemos que los análisis estadísticos, como los que se han propuesto, conducen a relaciones empíricas interesantes que pueden mejorar el desarrollo y evaluación de los sistemas de diálogo. Además, estos estudios, pueden servir para evaluar de forma global los sistemas en lugar de evaluar únicamente sus componentes individuales. Otra línea de trabajo futuro se centrará en desarrollar estudios de análisis factorial para agrupar criterios, obteniendo los aspectos principales que deben considerarse para optimizar el funcionamiento y maximizar la satisfacción del usuario. Con este fin, se generará una lista más extensa de criterios y una vez que se hayan evaluado los factores, se obtendrán las dependencias entre ellos para construir una taxonomía de criterios que poder comparar con otras del estado del arte como el *Quality-Of-Service* propuesto por Möller (2002).

Además, se introducirán y estudiarán nuevos criterios que permitan evaluar la inteligencia afectiva de los sistemas. De esta manera, será posible evaluar objetiva (por ejemplo, con respecto al éxito de la tarea) y subjetivamente (considerando las opiniones de los usuarios) los beneficios de integrar el modelo propuesto para el reconocimiento de emociones en el sistema de diálogo UAH (sección 6.2.1).

A
Publicaciones

La investigación descrita en la tesis se ha publicado en diversas actas de congresos nacionales e internacionales y revistas de impacto.

Se han realizado varias publicaciones describiendo las diferentes etapas de diseño e implementación del sistema UAH (capítulo 2). El siguiente artículo presenta el sistema de diálogo oral UAH, describiendo todos sus módulos y los enfoques innovadores empleados para su desarrollo.

- (Callejas y López-Cózar, 2005b) *Callejas, Z., López-Cózar, R., 2005. Implementing modular dialogue systems: a case study. En: Proc. of Applied Spoken Language Interaction in Distributed Environments (ASIDE 05). Aalborg, Denmark.*

Los artículos que se presentan a continuación se corresponden con una descripción y una demostración del módulo GAG del sistema UAH. El primero describe la versión inicial de dicho módulo, en la cual las reglas de las gramáticas se diseñaban y generaban automáticamente empleando una interfaz web, pero la actualización automática con los cambios en la base de datos todavía no había sido desarrollada.

- (Callejas y López-Cózar, 2005c) *Callejas, Z., López-Cózar, R., 2005. Nueva técnica de generación automática de gramáticas para sistemas de diálogo. Procesamiento del Lenguaje Natural (35), 205 - 212.*

- (Callejas y López-Cózar, 2005a) *Callejas, Z., López-Cózar, R., 2005. GAG: Generación automática de gramáticas en un sistema conversacional de interacción oral. Procesamiento del Lenguaje Natural (35), 457 - 458.*

La versión final de la herramienta GAG, en la que el proceso de creación y actualización de las gramáticas había sido automatizado por completo, se presenta en el siguiente artículo, donde se describe y evalúa con más detalle la técnica TGC.

- (Callejas y López-Cózar, 2007a) *Callejas, Z., López-Cózar, R., 2007. Automatic creation of ASR grammar rules for unknown vocabulary applications. En: Proc. of 8th International Workshop on Electronics, Control, Modelling, Measurement and Signals (ECMS). Liberec, Czech Republic. pp. 50-55.*

En cuanto a la investigación acerca del reconocimiento de emociones descrita en el capítulo 3, la siguiente publicación presenta un estado del arte en el campo del reconocimiento de emociones, prestando especial atención a los últimos proyectos internacionales y poniendo en contexto nuestra tarea investigadora.

- (Callejas y López-Cózar, 2007c) *Callejas, Z., López-Cózar, R., 2007. Emotion recognition for spoken dialogue systems. En: Proc. of I Simposio en Desarrollo de Software (SDS 2007). Granada, Spain. pp. 59-68.*

En el siguiente artículo se presenta un estudio completo acerca de los coeficientes Kappa y cómo se ven afectados por el desequilibrio de los corpus de emociones no actuadas.

- (Callejas y López-Cózar, 2008b) *Callejas, Z., López-Cózar, R., 2008. On the use of kappa coefficients to measure the reliability of the annotation of non-acted emotions. En: Proc. of 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems (PIT08). Kloster Irsee, Germany. Aceptado, será publicado en LNCS.*

La publicación que se muestra a continuación presenta la primera fase de desarrollo del enfoque de reconocimiento automático de emociones empleando información contextual. Describe la versión preliminar del algoritmo en dos etapas (sección 3.5).

- (Callejas y López-Cózar, 2007b) *Callejas, Z., López-Cózar, R., 2007. Decisive factors in the annotation of emotions for spoken dialogue systems. Advances in Soft Computing (45), 747 - 754.*

Los principales resultados experimentales del trabajo descrito en el capítulo 3 se presentan en el siguiente artículo, que detalla el trabajo realizado para añadir información contextual tanto para la anotación humana como para el aprendizaje automático. Incluye la versión final del algoritmo en dos etapas.

- (Callejas y López-Cózar, 2008a) *Callejas, Z., López-Cózar, R., 2008. Influence of contextual information in emotion annotation for spoken dialogue systems. Speech Communication Vol. 50 (5), 416-433.*

En cuanto a la investigación acerca de la adaptación entre idiomas del reconocedor de voz checo descrita en el capítulo 4, el siguiente artículo se corresponde con una demostración de la versión española del sistema MyVoice al reconocer español empleando los modelos acústicos del checo.

- (Callejas et al., 2007) *Callejas, Z., Nouza, J., Cerva, P., López-Cózar, R., 2007. Myvoice goes spanish. Cross-lingual adaptation of a voice-controlled PC tool for handicapped people. Procesamiento del Lenguaje Natural (39), 277 - 278.*

Finalmente, la investigación realizada para la evaluación de campo de los sistemas de diálogo (capítulo 5) se presenta en el siguiente trabajo.

- (Callejas y López-Cózar, 2008c) *Callejas, Z., López-Cózar, R. Relations between de-facto criteria in the evaluation of a spoken dialogue system. Speech Communication. Aceptado, disponible online desde el 15 de abril de 2008. DOI: 10.1016/j.specom.2008.04.004*

Bibliografía

Adell, J., Bonafonte, A., Escudero, D., 2005. Analysis of prosodic features towards modelling of emotional and pragmatic attributes of speech. Procesamiento de Lenguaje Natural 35, 277–284.

Ai, H., Litman, D., Forbes-Riley, K., Rotaru, M., Tetreault, J., Purandare, A., 2006. Using systems and user performance features to improve emotion detection in spoken tutoring dialogs. En: Proc. of the 9th Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pg. 797–800.

Alexandersson, J., Becker, T., 2001. Overlay as the basic operation for discourse processing in a multimodal dialogue system. En: Proc. of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems. Seattle, USA.

Andeani, G., Fabbrizio, D. D., Gilbert, M., Gillick, D., Hakkani-Tur, D., Lemon, O., 2006. Let's DISCOH: Collecting an Annotated Open Corpus with Dialogue Acts and Reward Signals for Natural Language Helpdesks. En: Proc. of IEEE 2006 Workshop on Spoken Language Technology (SLT'06). Palm Beach, Aruba, pg. 218–221.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. En: Proc. of the 7th International Conference on Spoken Language Processing (Interspeech'02-ICSLP). Denver, USA, pg. 2037–2040.

Artstein, R., Poesio, M., 2005. $kappa_3$ = alpha (or beta). Informe técnico, University of Essex.

AUBADE, 2005. `http://www.aubade-group.com/`.

Augmented Multiparty Interaction Project, 2007. `http://www.amiproject.org/`.

Baggia, P., Castagneri, G., Danieli, M., 2000. Field trials of the Italian ARISE train timetable system. Speech Communication 31, 355–367.

Bangalore, S., Hakkani-Tur, D., Tur, G., 2006. Introduction to the Special Issue on Spoken Language Understanding in Conversational Systems. Speech Communication 48, 233–3238.

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russel, M., Wong, M., 2004. Towards multilingual speech recognition using data driven source/target acoustical units association. En: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04). Montreal, Quebec, Canada, pg. 521–524.

Baudoin, F., Bretier, P., Corruble, V., 2005. A dialogue agent with adaptive and proactive capabilities. En: Proc. of IEEE/WIC/ACM International Conference on Intelligent Agent Technology. Compiegne, France, pg. 293–296.

Bayeh, R., Lin, S., Chollet, G., Mokbel, C., 2004. You stupid tin box! - children interacting with the aibo robot: a cross-linguistic emotional speech corpus. En: Proc. of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal, pg. 171–174.

Becker, T., Gerstenberger, C., Kruijff-Korbayova, I., Korthauer, A., Pinkal, M., Pitz, M., Poller, P., Schehl, J., 2006. Natural and intuitive multimodal dialogue for In-Car Applications: The SAMMIE System. En: Proc. of the 4th European Conference of Prestigious Applications of Intelligent Systems (PAIS'06). Riva del Garda, Italy, pg. 612–616.

Beringer, N., Kartal, U., Louka, K., Schiel, F., Tük, U., 2002. PROMISE: A Procedure for Multimodal Interactive System Evaluation. En: Proc. of the Workshop on Multimodal Resources and Multimodal Systems Evaluation. Las Palmas de Gran Canaria, Spain, pg. 77–80.

Bernsen, N., Dybkjaer, L., 1997. The DISC project. En: ELRA Newsletter. Vol. 2(2). pg. 6–8.

Bernsen, N., Dybkjaer, L., Dybkjaer, H., 1994. A dedicated task-oriented dialogue theory in support of spoken language dialogue system design. En:

Proc. of the 3rd International Conference on Spoken Language Processing (ICSLP'94). Yokohama (Japan), pg. 875–878.

Bernsen, N. O., Dybkjaer, L., 2000. A methodology for evaluating spoken language dialogue systems and their components. En: Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC'00). Athens, Greece, pg. 183–188.

Bickmore, T., Giorgino, T., 2004. Some novel aspects of health communication from a dialogue systems perspective. En: Proc. of AAAI Fall Symposium on Dialogue Systems for Health Communication. Washington DC, USA, pg. 275–291.

Bickmore, T., Giorgino, T., 2006. Health dialog systems for patients and consumers. Journal of Biomedical Informatics 39, 556–571.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Springer.

Boehner, K., DePaula, R., Dourish, P., Sengers, P., 2007. How emotion is made and measured. International Journal of Human-Computer Studies, Special Issue on Evaluating Affective Interactions 65 (4), 275–291.

Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Informe técnico, Institute of Phonetic Sciences, University of Amnsterdam.

Bohus, D., Grau, S., Huggins-Daines, D., Keri, V., Krishna, G., Kumar, R., Raux, A., Tomko, S., 2007. Conquest - an Open-Source Dialog System for Conferences. En: Proc. of Human Language Technologies'07: The Conference of the North American Chapter of the Association for Computational Linguistics. Rochester, NY, USA, pg. 9–12.

Bohus, D., Rudnicky, A., 2002. LARRI: A Language-Based Maintenance and Repair Assistant. En: Proc. of Multi-Modal Dialogue in Mobile Environments Conference (IDS'02). Kloster Irsee, Germany, pg. 203–218.

Bonaventura, P., Gallocchio, F., Micca, G., 1997. Multilingual speech recognition for flexible vocabularies. En: Proc. of 5th European Conference on Speech Communication and Technology (Eurospeech 1997). Rhodes, Greece, pg. 355–358.

Bos, J., Klein, E., Lemon, O., Oka, T., 1999. The verbmobil prototype system - a software engineering perspective. Journal of Natural Language Engineering 5(1), 95–112.

Boves, L., Os, E. D., 2002. Multimodal services, a MUST for UMTS. Informe técnico, EURESCOM.

Brown, M. K., 1999. Grammar Representation Requirements for Voice Markup Languages. Informe técnico, W3C.

Burkhardt, F., van Ballegooy, M., Englert, R., Huber, R., 2005. An emotion-aware voice portal. En: Proc. of Electronic Speech Signal Processing. Prague, Czech Republic, pg. 123–131.

Cahill, L., Tiberius, C., 2002. Cross-linguistic phoneme correspondences. En: Proc. of 19th International Conference on Computational Linguistics (COLING'02). Taipei, Taiwan, pg. 1–5.

Callas - Conveying Affectiveness in Leading-Edge Living Adaptive Systems, 2007. http://www.callas-newmedia.eu/.

Callejas, Z., López-Cózar, R., 2005a. GAG: Generación automática de gramáticas en un sistema conversacional de interacción oral. Procesamiento del Lenguaje Natural 35, 457–458.

Callejas, Z., López-Cózar, R., 2005b. Implementing modular dialogue systems: a case study. En: Proc. of Applied Spoken Language Interaction in Distributed Environments (ASIDE'05). Aalborg, Denmark.

Callejas, Z., López-Cózar, R., 2005c. Nueva técnica de generación automática de gramáticas para sistemas de diálogo. Procesamiento del Lenguaje Natural 35, 205–212.

Callejas, Z., López-Cózar, R., 2007a. Automatic creation of ASR grammar rules for unknown vocabulary applications. En: Proc. of the 8th International workshop on Electronics, Control, Modelling, Measurement and Signals (ECMS'07). Liberec, Czech Republic, pg. 50–55.

Callejas, Z., López-Cózar, R., 2007b. Decisive factors in the annotation of emotions for spoken dialogue systems. Advances in Soft Computing 45, 747–754.

Callejas, Z., López-Cózar, R., 2007c. Emotion recognition for spoken dialogue systems. En: Proc. of I Simposio en Desarrollo de Software (SDS'07). Granada, Spain, pg. 59–68.

Callejas, Z., López-Cózar, R., 2008a. Influence of contextual information in emotion annotation for spoken dialogue systems. Speech Communication 50 (5), 416–433.

Callejas, Z., López-Cózar, R., 2008b. On the use of kappa coefficients to measure the reliability of the annotation of non-acted emotions. En: Proc. of 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems (PIT'08). Kloster Irsee, Germany, to be published in LNCS.

Callejas, Z., López-Cózar, R., 2008c. Relations between de-facto criteria in the evaluation of a spoken dialogue system. Speech Communication In Press. Available online since 15th April 2008. DOI: 10.1016/j.specom.2008.04.004.

Callejas, Z., Nouza, J., Cerva, P., López-Cózar, R., 2007. Myvoice goes spanish. cross-lingual adaptation of a voice-controlled pc tool for handicapped people. Procesamiento del Lenguaje Natural 39, 277–278.

Camurri, A., Mazzarino, B., Volpe, G., 2004. Expressive interfaces. Cognition, Technology and Work 6 (1), 15–22.

Carreiras, M., García-Albea, J. E., Sebastián-Gallés, N., 1996. Language processing in Spanish. Lawrence Erlbaum Associates.

Castagneri, G., Baggia, P., Danieli, M., 1998. Field trials of the Italian ARISE train timetable system. En: Proc. of the Interactive Voice Technology for Telecommunications Applications Workshop (IVTTA'98). pg. 97–102.

Catizone, R., Setzer, A., Wilks, Y., 2003. Multimodal Dialogue Management in the COMIC Project. En: Proc. of EACL'03 Workshop on Dialogue Systems: interaction, adaptation, and styles of management. Budapest, Hungary, pg. 25–34.

Cerrato, L., 2002. A comparison between feedback strategies in human-to-human and human-machine communication. En: Proc. of the 8th International Conference on Spoken Language Processing (ICSLP 2002). Vol. 2. Denver, USA, pg. 557–560.

Cerva, P., Nouza, J., 2007. Design and development of voice controlled aids for motor-handicapped persons. En: Proc. of the 11th International Conference on Spoken Language Processing (Interspeech'07-Eurospeech). Antwerp, Belgium, pg. 2521–2524.

Chambers, N., Allen, J., 2004. Stochastic language generation in a dialogue system: Toward a domain independent generator. En: Proc. of the 5th SIGdial Workshop on Discourse and Dialogue. Boston, USA, pg. 9–18.

Cheyer, A., Julia, L., 1995. Multimodal maps: An agent based approach. En: Proc. of International Conference on Cooperative Multimodal Cooperation. Eindhoven, Holland, pg. 111–121.

CHIL - Computers In the Human Interaction Loop, 2007. `http://chil.server.de/servlet/is/101/`.

Chu, S.-W., O'Neill, I., Hanna, P., McTear, M., 2005. An approach to multi-strategy dialogue management. En: Proc. of 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech). Lisbon, Portugal, pg. 865–868.

Cicchetti, D. V., Feinstein, A. R., 1990. High agreement but low Kappa: II. Resolving the paradoxes. Journal of Clinical Epidemiology 43 (6), 551–558.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20 (3), 37–46.

Cohen, J., 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70 (4), 213–220.

Cole, R., Hirschman, L., Atlas, L., Beckman, M., Biermann, A., Bush, M., Clements, M., Cohen, J., Garcia, O., Hanson, B., Hermansky, H., Levinson, S., McKeown, K., Morgan, N., Novick, D. G., Ostendorf, M., Oviatt,

S., Price, P., Silverman, H., Spitz, J., Waibel, A., Weinstein, C., Zahorian, S., Zue, V., 1995. The Challenge of Spoken Language Systems: Research Direction for the Nineties. IEEE Transactions on Speech and Audio Processing 3, 1–20.

Cole, R., Mariani, J., Uszkoreit, H., Varile, G. B., Zaenen, A., Zampolli, A., Zue, V. (Eds.), 1997. Survey of the State of the Art in Human Language Technology. Cambridge University Press.

Companions, 2007. `http://nlp.shef.ac.uk/companions/`.

Corradini, A., Bernsen, N., Fredriksson, M., Johanneson, L., Königsmann, J., Mehta, M., 2004. Towards believable behavior generation for embodied conversational agents. En: Proc. of the Workshop on Interactive Visualisation and Interaction Technologies (IV&IT 2004). Krakow, Poland, pg. 946–953.

Corradini, A., Mehta, M., Bernsen, N. O., Charfuelán, M., 2005. Animating an interactive conversational character for an educational game system. En: Proc. of the 2005 International Conference on Intelligent User Interfaces. San Diego, CA, USA, pg. 183–190.

CoSy Home, 2007. `http://www.cs.bham.ac.uk/research/projects/cosy/`.

Cowie, R., 2000. Describing the emotional states expressed in speech. En: Proc. of the International Speech Communication Association (ISCA) Workshop on Speech and Emotion. Newcastle, Northern Ireland, UK, pg. 11–18.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M., 2000. FEELTRACE: An Instrument for Recording Perceived Emotion in Real Time. En: Proc. of the International Speech Communication Association (ISCA) Workshop on Speech and Emotion. Newcastle, Northern Ireland, UK, pg. 19–24.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., 2001. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18 (1), 32–80.

Cowie, R., Schröder, M., 2005. Piecing Together the Emotion Jigsaw. Lecture Notes on Computer Science 3361/2005, 305–317.

Craggs, R., Wood, M. M., 2003. Annotating emotion in dialogue. En: Proc. of the 4th SIGdial Workshop on Discourse and Dialogue. Sapporo, Japan, pg. 218–225.

Critchley, H. D., Rotshtein, P., Nagai, Y., ODoherty, J., Mathias, C. J., Dolana, R. J., 2005. Activity in the human brain predicting differential heart rate responses to emotional facial expressions. NeuroImage 24, 751–762.

Cuayáhuitl, H., Renals, S., Lemon, O., Shimodaira, H., 2006. Reinforcement learning of dialogue strategies with hierarchical abstract machines. En: Proc. of IEEE/ACL Workshop on Spoken Language Technology (SLT'06). Palm Beach, Aruba, pg. 182–186.

Dale, R., September 2003. Next-generation spoken dialog systems. Technology Trends Seminar Series, Macquarie University, http://www.ict.csiro.au/MU/Trends/2003.htm.

DARPA, 1992. Speech and Natural Language Workshop. Defense Advanced Research Projects Agency (DARPA), San Mateo, USA.

DARPA, 1994. Speech and Natural Language Workshop. Defense Advanced Research Projects Agency (DARPA), San Mateo, USA.

Davies, M., Fleiss, J. L., 1982. Measuring agreement for multinomial data. Biometrics 38 (4), 1047–1051.

de Melo, C., Paiva, A., 2005. Environment expression: Expressing emotions through cameras, lights and music. En: Proc. of the 1st International Conference on Affective Computing and Intelligent Interaction (ACII'05). Beijing, China, pg. 715–722.

Degerstedt, L., Jönsson, A., 2006. LinTest, A development tool for testing dialogue systems. En: Proc. of the 9th Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pg. 489–492.

den Os, E., Boves, L., Lamel, L., Baggia, P., 1999. Overview of the ARISE project. En: Proc. of the European Conference on Speech Technology (Eurospeech'99). Budapest, Hungary, pg. 1527–1530.

Devillers, L., Maynard, H., Rosset, S., 2004. The French Media/Evalda project: the evaluation of the understanding capability of spoken language dialog systems. En: Proc. of the 4th International Conference on Language Resources and Evaluation (LREC 2004). Vol. 6. Lisbon, Portugal, pg. 2131–2134.

Devillers, L., Vidrascu, L., Lamel, L., 2005. Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18, 407–422.

DISC, 1999. DISC Final Report covering the period from 1.6.98 to 28.2.99. Deliverable D5.2. Informe técnico, The DISC Consortium.

Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P., 2003. Emotional speech: towards a new generation of databases. Speech Communication 40, 33–60.

Dunn, G., 1989. Design and analysis of reliability studies: the statistical evaluation of measurement errors. Edward Arnold.

Dybkjaer, L., Bernsen, N. O., 2000. Usability issues in spoken language dialogue systems. Natural Language Engineering 6, 243–271.

Dybkjaer, L., Bernsen, N. O., Minker, W., 2004. Evaluation and usability of multimodal spoken language dialogue systems. Speech Communication 43, 33–54.

EAGLES, 1996. Evaluation of Natural Language Processing Systems. Final report. Document EAG-EWG-PR2. Informe técnico, Center for Sprogetknologi, Copenhague, Denmark.

EUROPA - CORDIS: Community Research and Development Information Service, 2006. `http://cordis.europa.eu/`.

Farzanfar, R., Frishkopf, S., Migneault, J., Friedman, R., 2004. Telephone-linked care for physical activity: A qualitative evaluation of the use patterns

of an information technology program for patients. Journal of Biomedical Informatics 38 (3), 220–228.

Feinstein, A. R., Cicchetti, D. V., 1990. High agreement but low Kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology 43 (6), 543–549.

Fleiss, J. L., 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin 76 (5), 378–382.

Fleiss, J. L., Cohen, J., 1973. The equivalence of weighted kappa and the interclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 33, 613–619.

Fluhr, C., Schmit, D., Andrieux, C., Ortet, P., Bisson, F., Combet, V., 1999. Crosslingual interrogation of multilingual catalogs. Lecture Notes on Computer Science 1696, 294–310.

Forbes-Riley, K., Litman, D. J., 2004a. Predicting emotion in spoken dialogue from multiple knowledge sources. En: Proc. of the Human Language Technology Conference - North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL'04). Boston, USA, pg. 201–208.

Forbes-Riley, K. M., Litman, D., 2004b. Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters. En: Proc. of the Human Language Technology Conference - North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL'06). New York, USA, pg. 264–271.

Gales, M. J. F., Woodland, P. C., 1996. Mean and Variance Adaptation Within the MLLR Framework. Computer Speech and Language 10, 249–264.

Gao, Y., Gu, L., Jeff, H.-K., 2005. Portability challenges in developing interactive dialogue systems. En: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05). Vol. 5. Philadelphia, PA, USA, pg. 18–23.

Gauvain, J., 1999. The limsi sdr system for trec. En: Proc. of the 8th Text Retrieval Conference (TREC 8). Gaithersburg, Maryland, USA, pg. 475–482.

Gauvain, J. L., Lee, C. H., 1994. Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. IEEE Transactions on Speech and Audio Processing 2, 291–298.

Gebhard, P., Klesen, M., Rist, T., 2004. Coloring multi-character conversations through the expression of emotions. En: Proc. of Tutorial and Research Workshop on Affective Dialogue Systems. Kloster Irsee, Germany, pg. 128–141.

Gerfen, C., 2002. Andalusian codas. Probus 14, 247–277.

Geutner, P., Steffens, F., Manstetten, D., 2002. Design of the VICO Spoken Dialogue System: Evaluation of User Expectations by Wizard-of-Oz experiments. En: Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC'02). Las Palmas de Gran Canaria, Spain, pg. 385–400.

Gibbon, D., Mertins, I., Moore, R., 2000. Handbook of Multimodal and Spoken Dialogue Systems. Resources, Terminology and Product Evaluation. Kluwer Academic Publishers (Kluwer International Series in Engineering and Computer Science, 565).

Gibbon, D., Moore, R., Winski, R., 1997. Handbook of Standards and Resources for Spoken Language Systems. Walter de Gruyter.

Glass, J., Flammia, G., Goodine, D., Phillips, M., Polifroni, J., Sakai, S., Seneff, S., Zue, V., 1995. Multilingual spoken-language understanding in the MIT Voyager system. En: Speech Communication. Vol. 17. pg. 1–18.

Glass, J. R., 1999. Challenges for spoken dialogue systems. En: Proc. of the Automatic Speech Recognition and Understanding Workshop (ASRU'99). Keystone, Colorado, USA.

González, G. M., 1999. Bilingual computer-assisted psychological assessment: An innovative approach for screening depression in chicanos/latinos. Informe Técnico 39, University of Michigan, Ann Arbor, USA.

Griol, D., 2007. Desarrollo y Evaluación de diferentes Metodologías para la Gestión Automática del Diálogo. Tesis Doctoral, Universidad Politécnica de Valencia, Valencia, Spain.

Gupta, N., Tur, G., Hakkani-Tur, D., Bangalore, S., Riccardi, G., Gilbert, M., 2006. The AT&T spoken language understanding system. IEEE Transactions on Audio, Speech and Language processing 14, 213–222.

Gustafson, J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D., Wirén, M., 2000. AdApt - a multimodal conversational dialogue system in an apartment domain. En: Proc. of the 6th International Conference on Spoken Language Processing (ICSLP'00). Vol. 2. Beijing, China, pg. 134–137.

Gut, U., Bayerl, P. S., 2004. Measuring the reliability of manual annotations of speech corpora. En: Proc. of the 2nd International Conference on Speech Prosody (SP'04). Nara, Japan, pg. 565–568.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182.

Haas, J., Gallwitz, F., Horndasch, A., Huber, R., Warnke, V., 2005. Telephone-based speech dialog systems. Lecture Notes on Computer Science 3663, 125–132.

Hall, L., Woods, S., Aylett, R., Paiva, A., Newall, L., 2005. Achieving empathic engagement through affective interaction with synthetic characters. En: Proc. of the 1st International Conference on Affective Computing and Intelligent Interaction (ACII'05). Beijing, China, pg. 731–738.

Hamerich, S., de Córdoba, R., Schless, V., d'Haro, L., Schubert, V., Kocsis, O., Igel, S., Pardo, J. M., 2004. The GEMINI Platform: Semi-Automatic Generation of Dialogue Applications. En: Proc. of the 8th International Conference on Spoken Language Processing (Interspeech-ICSLP). Vol. 4. Jeju, Corea, pg. 2629–2632.

Hanna, P., O'Neill, I., Wootton, C., McTear, M., 2007. Promoting extension and reuse in a spoken dialog manager: an evaluation of the Queen's Communicator. ACM Transactions on Speech and Language Processing 4.

Hansen, J. H. L., 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Speech communication 20 (2), 151–170.

Hardy, H., Biermann, A., Inouye, R., McKenzie, A., Strzalkowski, T., Ursu, C., Webb, N., Wu, M., 2006. The Amitiés system: Data-driven techniques for automated dialogue. Speech Communication 48, 354–373.

Hartikainen, M., Salonen, E.-P., Turunen, M., 2004. Subjective Evaluation of Spoken Dialogue Systems Using SERVQUAL Method. En: Proc. of 8th International Conference on Spoken Language Processing (Interspeech'04-ICSLP). Jeju Island, Korea, pg. 2273–2276.

Haseel, L., Hagen, E., 2005. Adaptation of an automotive dialogue system to users' expertise. En: Proc. of 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech). Lisbon, Portugal, pg. 222–226.

Heim, J., Nilsson, E. G., Skjetne, J. H., 2007. User Profiles for Adapting Speech Support in the Opera Web Browser to Disabled Users. Lecture Notes on Computer Science 4397, 154–172.

Higashinaka, R., Miyazaki, N., Nakano, M., Aikawa, K., 2004. Evaluating discourse understanding in spoken dialogue systems. ACM Transactions on Speech and Language Processing 1, 1–20.

Holmes, W., Huckvale, M., 1994. Why have HMMs been so successful for automatic speech recognition and how might they be improved? En: Speech Hearing and Language. pg. 1875–1878.

Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., Nogueiras, A., 2002. Interface databases: Design and collection of a multilingual emotional speech database. En: Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC'02). Las Palmas de Gran Canaria, Spain, pg. 385–400.

Hu, Y., Loizou, P. C., 2007. Subjective comparison and evaluation of speech enhacement algorithms. Speech Communication 49, 588–601.

Hubal, R. C., Frank, G. A., Guinn, C. I., 2000. AVATALK Virtual Humans for Training with Computer Generated Forces. En: Proc. of the 9th Conference on Computer Generated Forces and Behavioral Representation. Orlando, Florida, USA, pg. 617–623.

Humaine emotion-research.net, 2007. http://emotion-research.net/.

Hurtig, T., 2004. Visualization and multimodality: a mobile multimodal dialogue system for public transportation navigation evaluated. En: Proc. of the 8th Conference on Human-computer interaction with mobile devices and services (MobileHCI'06). Helsinki, Finland, pg. 251–254.

Interspeech 2007 Special Session on Speech and language technology for less-resourced languages, 2007. http://www.interspeech2007.org/Technical/less_resourced_languages.php.

Intrepid Project (A virtual reality intelligent multi-sensor wearable system for phobias' treatment), 2004. http://www.intrepid-project.org/.

Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J. M., Bernadas, D., Oliver, J. M., Tena, D., Longhi, L., 2000. Validation of an acoustical modelling of emotional expression in spanish using speech synthesis techniques. En: Proc. of the International Speech Communication Association (ISCA) Workshop on Speech and Emotion. Newcastle, Northern Ireland, UK, pg. 161–166.

J. Gustafson, N. L., Lundeberg, M., 1999. The August spoken dialogue system. En: Proc. of the 6th European Conference on Speech Communication and Technology (EuroSpeech'99). Budapest, Hungary, pg. 1151–1154.

Johnstone, T., 1996. Emotional speech elicited using computer games. En: Proc. of the 4th International Conference on Spoken Language Processing (ICSLP 1996). Vol. 3. Philadelphia, PA, pg. 1985–1988.

Jokinen, K., 2003. Natural interaction in spoken dialogue systems. En: Proc. of the Workshop Ontologies and Multilinguality in User Interfaces. Crete, Greece, pg. 730–734.

Jokinen, K., Kanto, K., Rissanen, J., 2004. Adaptative User Modelling in AthosMail. Lecture Notes on Computer Science 3196, 149–158.

Jurafsky, D., Martin, J. H., 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall.

Kacic, Z., 1999. Advances in spoken dialogue systems development. En: Proc. of IEEE International Symposium on Industrial Electronics (ISIE'99). Vol. 1. Bled, Slovenia, pg. 169–172.

Kirchhoff, K., Bilmes, J. A., 1999. Statistical acoustic indications of coarticulation. En: Proc. of International Congress of Phonetic Sciences. San Francisco, California, USA, pg. 1729–1732.

Kirchhoff, K., Vergyri, D., 2005. Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition. Speech Communication 46, 37–51.

Krippendorff, K., 2003. Content Analysis: An Introduction to its Methodology. Sage Publications, Inc.

Krsmanovic, F., Spencer, C., Jurafsky, D., Ng, A. Y., 2006. Have we meet? MDP Based Speaker ID for Robot Dialogue. En: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pg. 461–464.

Kumar, C. S., Mohandas, V. P., Haizhou, L., 2005. Multilingual speech recognition: A unified approach. En: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech). Lisbon, Portugal, pg. 3357–3360.

Kwon, O., Yoo, K., Suh, E., 2005. ubiES: An Intelligent Expert System for Proactive Services Deploying Ubiquitous Computing Technologies. En: Proc. of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05). Hawaii, pg. 85–86.

Lamel, L., Bennacef, S., Gauvain, J., Dartigues, H., Temem, J., 2002. User evaluation of the MASK kiosk. Speech Communication 38 (1-2), 131–139.

Lamel, L., Minker, W., Paroubek, P., 2000a. Towards best practice in the development and evaluation of speech recognition components of a spoken language dialog system. Natural Language Engineering 6 (3-4), 305–322.

Lamel, L., Rosset, S., Gauvain, J., Bennacef, S., Garnier-Rizet, M., Prouts, B., 2000b. The LIMSI ARISE system. Speech Communication 31, 339–353.

Landis, J. R., Koch, G. G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Langner, B., Black, A., 2005. Using speech in noise to improve understandability for elderly listeners. En: Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'05). San Juan, Puerto Rico, pg. 392–396.

Lantz, C. A., Nebenzahl, E., 1996. Behavior and interpretation of the κ statistic: Resolution of the two paradoxes. Journal of Clinical Epidemiology 49 (4), 431–434.

Larsen, L. B., 2003. Issues in the evaluation of spoken dialogue systems using objective and subjective measures. En: Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03). St. Thomas, U.S. Virgin Islands, USA, pg. 209–214.

Lee, C., Yoo, S. K., Park, Y. J., Kim, N. H., Jeong, K. S., Lee, B. C., 2005. Using Neural Network to Recognize Human Emotions from Heart Rate Variability and Skin Resistance. En: Proc. of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'05). Shanghai, China, pg. 5523–5525.

Lee, C. M., Narayanan, S. S., 2005. Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing 13 (2), 293–303.

Legris, P., Ingham, J., Collerette, P., 2003. Why do people use information technology: A critical review of the technology acceptance model. Information and Management 40, 191–204.

Lehmann, S., Oepen, S., Regnier-Prost, S., Netter, K., Lux, V., Klein, J., Falkedal, K., Fouvry, F., Estival, D., Dauphin, E., Compagnion, H., Baur, J., Balkan, L., Arnold, D., 1996. TSNLP — Test Suites for Natural Language Processing. En: Proc. of the 16th International Conference on Computational Linguistics (COLING'96). Copenhagen, Denmark, pg. 711–716.

Lemon, O., Bracy, A., Gruenstein, A., Peters, S., 2001. The Witas Multi-Modal Dialogue System. En: Proc. of Eurospeech 2001. Aalborg, Denmark, pg. 1559–1562.

Lemon, O., Georgila, K., Henderson, J., 2006. Evaluating Effectiveness and Portability of Reinforcement Learned Dialogue Strategies with real users: the TALK TownInfo Evaluation. En: Proc. of IEEE-ACL Workshop on Spoken Language Technology (SLT 2006). Palm Beach, Aruba, pg. 178–181.

Ligozat, A.-L., Grau, B., Robba, I., Vilnat, A., 2006. Evaluation and improvement of cross-lingual question answering strategies. En: Proc. of the 11th EACL Workshop on Multilingual Question Answering (MLQA'06). Trento, Italy, pg. 23–30.

Lines, L., Hone, K. S., 2002. Older adults' evaluations of speech output. En: Proc. of the 5th International ACM SIGCAPH Conference on Assistive Technologies (ASSETS'02). Edinburgh, Scotland, pg. 170–177.

Liscombe, J., Riccardi, G., Hakkani-Tür, D., 2005. Using context to improve emotion detection in spoken dialog systems. En: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech). Lisbon, Portugal, pg. 1845–1848.

Litman, D. J., Forbes-Riley, K., 2006. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. Speech communication 48 (5), 559–590.

Litman, D. J., Pan, S., 2002. Designing and evaluating an adaptive spoken dialogue system. User modelling and user-adapted interaction 12, 111–137.

Litman, D. J., Silliman, S., 2004. ITSPOKE: An Intelligent Tutoring Spoken Dialogue System. En: Proc. of the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL). Boston, USA, pg. 233–236.

López-Cózar, R., Araki, M., 2005. Spoken, Multilingual and Multimodal Dialogue Systems. Development and Assessment. John Wiley and Sons.

LREC 2008 Workshop on Collaboration: interoperability between people in the creation of language resources for less-resourced languages, 2008. http://ixa2.si.ehu.es/saltmil/en/home/.

MagiCster Project Pages, 2007. http://www.ltg.ed.ac.uk/magicster/.

Mahlke, S., 2006. Emotions and EMG measures of facial muscles in interactive contexts. En: Proc. of the 2006 Conference on Human Factors in Computer Systems (CHI'06). Montreal, Canada.

Mangold, H., 2001. Speech technology in reality - Applications, their challenges and solutions. En: Proc. of the 4th International Conference on Text, Speech and Dialogue (TSD'01). Pilsen, Czech Republic, pg. 197–200.

Martinovski, B., Traum, D., 2003. Breakdown in human-machine interaction: the error is the clue. En: Proc. of the ISCA tutorial and research workshop on Error handling in dialogue systems. Chateau d'Oex, Vaud, Switzerland, pg. 11–16.

Martín-Valdivia, M. T., Martínez-Santiago, F., Ureña-López, L., 2005. Merging strategy for cross-lingual information retrieval systems based on learning vector quantization. Neural Processing Letters 22 (2), 149–161.

Mattasoni, M., Omologo, M., Santarelli, A., Svaizer, P., 2002. On the Joint Use of Noise Reduction and MLLR Adaptation for In-Car Hands-Free Speech Recognition. En: Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02). Orlando, USA, pg. 289–292.

McTear, M. F., 2004. Spoken dialogue technology. Springer.

MEGA Project, 2001. http://www.megaproject.org/.

Melin, H., Sandell, A., Ihse, M., 2001. Ctt-bank: A speech controlled telephone banking system - an initial evaluation. En: TMH-QPSR. Vol. 1. pg. 1–27.

Menezes, P., Lerasle, F., Dias, J., Germa, T., 2007. Towards an interactive humanoid companion with visual tracking modalities. International Journal of Advanced Robotic Systems, 48–78.

Mihalcea, R., Leong, B., 2006. Toward communicating simple sentences using pictorial representations. En: Proc. of the 7th Conference of the Association for Machine Translation in the Americas (AMTA'06). Boston, MA, USA, pg. 119–127.

Minker, W., 1998. Stochastic versus rule-based speech understanding for information retrieval. Speech Communication 25(4), 223–247.

Minker, W., Albalate, A., Bühler, D., Pittermann, A., Pittermann, J., Strauss, P.-M., Zaykovskiy, D., 2006a. Recent trends in spoken language dialogue systems. En: ITI 4th International Conference on Information and Communications (ICICT'06). Montreal, Canada, pg. 1–16, invited paper.

Minker, W., Haiber, U., Heisterkamp, P., Scheible, S., 2004a. The Seneca Spoken Language Dialogue System. En: Speech Communication. Vol. 43. pg. 1–2.

Minker, W., Haiber, U., Heisterkamp, P., Scheible, S., 2004b. The SENECA spoken language dialogue system. Speech Communication 43, 89–102.

Minker, W., Pittermann, J., Pittermann, A., Strauss, P.-M., Bühler, D., 2006b. Next-generation human-computer interfaces - Towards intelligent, adaptive and proactive spoken language dialogue systems. En: Proc. of the 2nd IET International Conference on Intelligent Environments (IE'06). Vol. 1. Athens, Greece, pg. 213–219.

Mäkelä, K., Salonen, E., M., T., Hakulinen, J., Raisamo, R., 2001. Evaluating the User Interface of a Ubiquitous Computing system Doorman. En: Proc. of the 3rd International Conference on Ubiquitous Computing (Ubicomp'01). Atlanta, USA.

Möller, S., 2002. A new taxonomy for the quality of telephone services based on spoken dialogue systems. En: Proc. of the 3rd Workshop on Discourse and Dialogue (SIGDial'02). Philadelphia, USA, pg. 142–153.

Möller, S., 2005. Quality of telephone-based spoken dialogue systems. Springer.

Möller, S., Smeele, P., Boland, H., Krebber, J., 2007. Evaluating spoken dialogue systems according to de-facto standards: A case study. Computer Speech and Language 21, 26–53.

Montero, J. M., Gutiérrez-Arriola, J., Enríquez, E., Pardo, J. M., 1999. Analysis and modelling of emotional speech in Spanish. En: Proc. of the 14th International Conference of Phonetic. San Francisco, USA, pg. 957–960.

Montoro, G., Alamán, X., Haya, P. A., 2004. Spoken interaction in intelligent environments: a working system. En: Ferscha, A., Hoertner, H., Kotsis, G. (Eds.), Advances in Pervasive Computing. Austrian Computer Society (OCG), pg. 747–754.

Montoro, G., Haya, P. A., Alamán, X., López-Cózar, R., Callejas, Z., 2006. A proposal for an XML definition of a dynamic spoken interface for ambient intelligence. En: International Conference on Intelligent Computing (ICIC'06). Kunming, China, pg. 711–716.

Morgan, N., Fosler, E., Mirghafori, N., 1997. Speech recognition using on-line estimation of speaking rate. En: Proc. of the 5th European Conference on Speech Communication and Technology (Eurospeech'97). Rhodes, Greece, pg. 2079–2082.

Morrison, D., Wang, R., Silva, L. C. D., 2007. Ensemble methods for spoken emotion recognition in call-centers. Speech communication 49, 98–112.

Mostow, J., 2008. Experience from a reading tutor that listens: Evaluation purposes, excuses, and methods. En: Kinzer, C., Verhoeven, L. (Eds.), Interactive Literacy Education: Facilitating Literacy Environments Through Technology. New York: Lawrence Erlbaum Associates, Taylor and Francis, pg. 117–148.

Mourão, M., Cassaca, R., Mamede, N., 2004. An Independent Domain Dialogue System Through a Service Manager. En: Proc. of the 4th International Confernce on Advances in Natural Language Processing. Alicante, Spain, pg. 161–171.

Nakamura, S., Markov, K., Jitsuhiro, T., Zhang, J.-S., Yamamoto, H., Kikui, G., 2004. Multi-lingual speech recognition system for speech-to-speech translation. En: Proc. of 8th International Conference on Spoken Language Processing (Interspeech'04-ICSLP). Jeju Island, Korea, pg. 146–154.

NECA project, 2005. `http://www.ofai.at/research/nlu/NECA`.

Nguyen, A., Wobcke, W., 2006. Extensibility and Reuse in an Agent-Based Dialogue Model. En: Proc. of IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. pg. 367–371.

NICE project - Main page, 2007. `http://www.niceproject.com/`.

Nielsen, J., 1994. Usability Engineering. Morgan Kaufmann Ed., San Francisco, USA.

Nieuwoudt, C., Botha, E. C., 1999. Cross-language adaptation of acoustic models in automatic speech recognition. En: Proc. of 5th Africon Conference in Africa (IEEE Africon 1999). Cape Town, South Africa, pg. 181–184.

Nieuwoudt, C., Botha, E. C., 2002. Cross-language use of acoustic information for automatic speech recognition. Speech Communication 38, 101–113.

Németh, G., Zainkó, C., 2003. Multilingual statistical text analysis, Zipf's law and Hungarian speech generation. Acta Linguistica Hungarica 49, 385–405.

Nouza, J., Nouza, T., Cerva, P., 2005. A multi-functional voice-control aid for disabled persons. En: Proc. of International Conference on Speech and Computer (SPECOM'05). Patras, Greece, pg. 715–718.

Nouza, J., Psutka, J., Uhlír, J., 1997. Phonetic Alphabet for Speech Recognition of Czech. Radioengineering 6 (4), 16–20.

Odel, J., Mukerjee, K., 2007. Architecture, user interface, and enabling technology in Windows Vista's speech systems. IEEE Transactions on Computers 56 (9), 1156–1168.

Oh, A., Rudnicky, A., 2000. Stochastic language generation for spoken dialogue systems. En: Proc. of ANLP/NAACL 2000 Workshop on Conversational Systems. Seattle, USA, pg. 27–32.

O'Neill, P., 2005. Utterance final /s/ in Andalusian Spanish. The phonetic neutralization of a phonological contrast. Language Design 7, 151–166.

Ortony, A., G. L. Clore, A. C., 1988. The cognitive structure of emotions. Cambridge University Press.

Ostendorf, M., Digalakis, V., Kimball, O., Sep 1996. From hmm's to segment models: a unified view of stochastic modeling for speech recognition. Speech and Audio Processing, IEEE Transactions on 4 (5), 360–378.

Oviatt, S., DeAngeli, A., Kuhn, K., 1997. Integration and synchronization of input modes during multimodal human-computer interaction. En: Proc. of the SIGCHI conference on Human factors in computing systems. Atlanta, Georgia, USA, pg. 415–422.

Paek, T., 2001. Empirical methods for evaluating dialog systems. En: Proc. of the Workshop on Evaluation for Language and Dialogue Systems. Vol. 9. Toulouse, France, pg. 1–8.

Park, W., Han, S. H., Park, Y. S., Park, J., Yang, H., 2007. A framework for evaluating the usability of spoken language dialogue systems (SLDSs). Lecture Notes on Computer Science 4559, 398–404.

Pérez, G., Amores, G., Manchón, P., 2006. A multimodal architecture for home control by disabled users. En: Proc. of IEEE/ACL Workshop on Spoken Language Technology (SLT). Palm Beach, Aruba, pg. 134–137.

Persia, L. D., Yanagida, M., Rufiner, H. L., Milone, D., 2007. Objective quality evaluation in blind source separation for speech recognition in a real room. Signal Processing 87, 1951–1963.

PF-STAR home page, 2004. `http://pfstar.itc.it/`.

Pfau, T., Ruske, G., 1998. Estimating the speaking rate by vowel detection. En: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98). Vol. 2. Seattle, Washington, USA, pg. 945–948.

Picard, R. W., 1997. Affective Computing. The MIT Press, Cambridge, Massachusetts.

Picard, R. W., Daily, S. B., 2005. Evaluating affective interactions: Alternatives to asking what users feel. En: Proc. of the 2005 Conference on Human Factors in Computer Systems (CHI'05), Workshop: Evaluating Affective Interfaces-Innovative Approaches. Portland, Oregon, USA.

Pieraccini, R., Levin, E., Eckert, W., 1997. AMICA: The AT&T mixed initiative conversational architecture. En: Proc. of European Conference on Speech Communications and Technology (Eurospeech'97). Rhodes, Greece, pg. 1875–1878.

Pitterman, J., Pitterman, A., 2006. Integrating emotion recognition into an adaptive spoken language dialogue system. En: Proc. of the 2nd IEEE International Conference on Intelligent Environments. Athens, Greece, pg. 197–202.

Plutchik, R., 1980. EMOTION: A psychoevolutionary synthesis. Harper and Row publishers.

Polzin, T., Waibel, A., 2000. Emotion-sensitive human-computer interfaces. En: Proc. of ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion. Newcastle, Northern Ireland, UK, pg. 201–206.

Prendinger, H., Mayer, S., Mori, J., Ishizuka, M., 2003. Persona effect revisited. using bio-signals to measure and reflect the impact of character-based interfaces. En: Proc. of the 4th International Working Conference on Intelligent Virtual Agents (IVA'03). Kloster Irsee, Germany, pg. 283–291.

Radev, D., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Celebi, A., Qi, H., Drabek, E., Liu, D., 2001. Evaluation of text summarization in a cross-lingual information retrieval framework. Informe técnico, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA.

Rajman, M., Bui, T. H., Rajman, A., Seydoux, F., Trutnev, A., Quarteroni, S., 2004. Assessing the usability of a dialogue management system designed in the framework of a rapid dialogue prototyping methodology. Acta Acustica united with Acustica 90, 1906–1111.

Raux, A., Bohus, D., Black, A. W., Eskenazi, M., 2006. Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go! Experience. En: Proc. of the 9th Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pg. 65–68.

Raux, A., Langner, B., Black, A., Eskenazi, M., 2005. Let's Go Public! Taking a Spoken Dialog System to the Real World. En: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech). Lisbon, Portugal, pg. 885–888.

Raux, A., Langner, B., Black, A. W., Eskenazi, M., 2003. LET'S GO: Improving Spoken Dialog Systems for the Elderly and Non-natives. En: Proc. of the European Conference on Speech Technology (Eurospeech'03). Geneva, Switzerland, pg. 753–756.

Rayner, M., Hockey, B., Renders, J.-M., Chatzichrisafis, N., Farrell, K., 2005. A voice enabled procedure browser for the International Space Station. En: 43th Annual Meeting of the Association for Computational Linguistics. Ann Arbor, USA, pg. 29–32.

Riccardi, G., Hakkani-Tür, D., 2005. Grounding emotions in human-machine conversational systems. Lecture Notes in Computer Science, 144–154.

Rich, C., Sidner, C., 1998. COLLAGEN: A Collaboration Manager for Software Interface Agents. User Modeling and User-Adapted Interaction 8, 315–350.

Robinson, S. M., Roque, A., Vaswani, A., Traum, D., 2006. Evaluation of a spoken dialogue system for virtual reality call for fire training. En: Proc. of the 25th Army Science Conference. Orlando, USA.

Roque, A., Ai, H., Traum, D., 2006a. Evaluation of an information state-based dialogue manager. En: Proc. of the 10th Workshop on the Semantics and Pragmatics of Dialogue (Brandial'06). Potsdam, Germany, pg. 181–182.

Roque, A., Leuski, A., Rangarajan, V., Robinson, S., Vaswani, A., Narayanan, S., Traum, D., 2006b. Radiobot-CFF: A Spoken Dialogue System for Military Training. En: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pg. 477–480.

Rotaru, M., Litman, D. J., 2006. Discourse structure and speech recognition problems. En: Proc. of the 9th Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pg. 53–56.

Rudnicky, A., Thayer, E., Constantinides, P., Tchou, C., Shern, R., Lenzo, K., Xu, W., Oh, A., 1999. Creating natural dialogs in the Carnegie Mellon Communicator system. En: Proc. of European Conference on Speech Communications and Technology (Eurospeech'99). Vol. 1(4). pg. 1531–1534.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., 1986. PDP: Computational models of cognition and perception, I. MIT Press.

Russell, J. A., 1980. A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178.

Sanderman, A., Sturm, J., den Os, E., Boves, L., Cremers, A., 1998. Evaluation of the Dutch train timetable information system developed in the ARISE project. En: Proc. of IEEE 4th Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA'98). Torino, Italy, pg. 91–96.

Sarukkai, R., Hunter, C., 1997. Integration of eye fixation information with speech recognition systems. En: Proc. of the 5th European Conference on Speech Communication and Technology (Eurospeech'97). Rhodes, Greece, pg. 1639–1643.

Schalkwyck, J., Story, L. H. E., 2003. Speech recognition with Dynamic Grammars Using Finite-State Transducer. En: Proc. of Eurospeech'03. Geneva, Switzerland, pg. 1969–1972.

Scherer, K. R., 2005. What are emotions? and how can they be measured? Social Science Information 44 (4), 694–729.

Schiel, F., 2006. Evaluation of multimodal dialogue systems. En: Wahlster (2006), pg. 617–643.

Schneider, M., 2004. Towards a Transparent Proactive User Interface for a Shopping Assistant. En: Proc. of Workshop on Multi-User and Ubiquitous User Interfaces (MU3I). Vol. 3. Funchal, Madeira, Portugal, pg. 10–15.

Schultz, T., Kirchhoff, K., 2006. Multilingual Speech Processing. Elsevier.

Schultz, T., Waibel, A., 1998. Language independent and language adaptive large vocabulary speech recognition. En: Proc. of the 5th International Conference of Spoken Language Processing (ICSLP'98). Vol. 5. Sidney, Australia, pg. 1819–1822.

Scott, W., 1955. Reliability of content analysis: the case of nominal scale coding. Public opinion quarterly 19 (3), 321–325.

Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., Zue, V., 1998. Galaxy-ll: A Reference Architecture for Conversational System Development. En: Proc. of the 5th International Conference on Spoken Language Processsing (ICSLP'98). Vol. 3. Sydney, Australia, pg. 931–934.

Seneff, S., Polifroni, J., 2000. Dialogue management in the Mercury flight reservation system. En: Proc. of ANLP-NAACL Workshop on Conversational systems. Vol. 3. Seattle, Washington, USA, pg. 11–16.

Shafran, I., Mohri, M., 2005. A comparison of classifiers for detecting emotion from speech. En: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05). Philadelphia, PA, USA, pg. 341–344.

Shafran, I., Riley, M., Mohri, M., 2003. Voice signatures. En: Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03). St. Thomas, U.S. Virgin Islands, USA, pg. 31–36.

Sim, J., Wright, C. C., 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy 85 (3), 257–268.

Skantze, G., Edlund, J., Carlson, R., 2006. Talking with Higgins: Research challenges in a spoken dialogue system. En: Proc. of Perception and Interactive Technologies (PIT'06). Kloster Irsee, Germany, pg. 193–196.

Stern, R., Liu, F., Ohshima, Y., Sullivan, T., Acero, A., 1992. Multiple approaches to robust speech recognition.

Stewart, J., 1922. An elecrical analog of the vocal tract. Nature 110, 311–312.

Stibbard, R., 2000. Automated extraction of tobi annotation data from the reading/leeds emotional speech corpus. En: Proc. of the International Speech Communication Association (ISCA) Workshop on Speech and Emotion. Newcastle, Northern Ireland, UK, pg. 60–65.

Sturm, J., Cranen, B., Terken, J., Bakx, I., 2005. Effects of prolonged use on the usability of a multimodal form-filling interface. En: Minker, W., Bühler, D., Dybkjaer, L. (Eds.), Spoken multimodal human-computer dialogue in mobile environments. Springer, pg. 329–348.

Sun, D., 1997. Statistical modeling of co-articulation in continuous speech based on data driven interpolation. En: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97). Munich, Germany, pg. 1751–1754.

The Safira Project - DFKI Page, 2002. http://www2.dfki.de/imedia/safira/.

Traum, D., Larsson, S., 2003. The Information State Approach to Dialogue Management. Current and New Directions in Discourse and Dialogue. Kluwer Academic Publishers.

TRINDIConsortium, 2001. TRINDI (Task Oriented Instructional Dialogue) Book Draft. http://www.ling.gu.se/projekt/trindi/book.ps .

Truillet, P., Grisvard, O., Goujon, B., 2004. SCOPE - CARE II Innovative WP3 -R3 - Model of English. Informe técnico, European Organisation for the Safety of Air Navigation.

Turing, A., 1950. Computing machinery and intelligence. Mind 236, 433–460.

Turunen, M., Hakulinen, J., Kainulainen, A., 2006. Evaluation of a spoken dialogue system with usability tests and long-term pilot studies: Similarities and differences. En: Proc. of the 9th Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pg. 1057–1060.

Turunen, M., Salonen, E., Hartikainen, M., Hakulinen, J., Black, W., Ramsay, A., Funk, A., Conroy, A., Thompson, P., Stairmand, M., Jokinen, K., Rissanen, J., Kanto, K., Kerminen, A., Gamback, B., Cheadle, M., Olsson, F., Sahlgren, M., 2004. Athosmail: a multilingual adaptive spoken dialogue system for the e-mail domain. En: Proc. of Workshop on Robust and

Adaptive Information Processing for Mobile Speech Interfaces. Geneva, Switzerland, pg. 77–86.

Vaquero, C., Saz, O., Lleida, E., Marcos, J., Canalís, C., 2006. VOCALIZA: An application for computer-aided speech therapy in spanish language. En: Proc. of IV Jornadas en Tecnología del Habla. Zaragoza, Spain, pg. 321–326.

Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: resources, features and methods. Speech communication 48, 1162–1181.

VICTEC in Lynne Hall web page, 2005. http://osiris.sunderland.ac.uk/~cs0lha/Research/victec.html.

Vidrascu, L., Devillers, L., 2005. Real-life emotion representation and detection in call centers data. Lecture Notes on Computer Science 3784, 739–746.

Villing, J., Larsson, S., 2006. Dico - A Multimodal Menu-based In-vehicle Dialogue System. En: Proc. of the 10th Workshop on the Semantics and Pragmatics of Dialogue (BRANDIAL'06). Potsdam, Germany, pg. 187–188.

Vogt, T., André, E., 2005. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. En: Proc. of IEEE International Conference on Multimedia and Expo. pg. 474–477.

Wahlster, W. (Ed.), 2006. SmartKom: Foundations of Multimodal Dialogue Systems. Springer.

Waibel, A., Suhm, B., Vo, M., Yang, J., 1997. Multimodal Interfaces for Multimedia Information Agents. En: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97). Munich, Germany, pg. 167–170.

Walker, M., Cahn, J., Whittaker, S., 1997. Improvising linguistic style: Social and affective bases of agent personality. En: Proc. of the 1st International Conference on Autonomous Agents (Agents'97). Marina del Rey, CA, USA, pg. 96–105.

Walker, M., Fromer, J., Fabbrizio, G., Mestel, C., Hindle, D., 1998a. What can I say? Evaluating a spoken language interface to Email. En: Proc. of ACM CHI 98 Conference on Human Factors in Computing Systems. Los Angeles, USA, pg. 582–589.

Walker, M., Kamm, C. A., Litman, D. J., 2000a. Towards developing general models of usability with paradise. Natural Language Engineering, 363–377.

Walker, M., Langkilde, I., wright, J., Gorin, A., Litman, D., 2000b. Learning to predict problematic situations in a spoken dialogue system: Experiments with How May I Help You. En: Proc. of the North American Meeting of the Association for Computational Linguistics. Seattle, USA, pg. 210–217.

Walker, M., Litman, D., Kamm, C., Abella, A., 1998b. Evaluating spoken dialogue agents with PARADISE: Two case studies. Computer Speech and Language 12, 317–347.

Walker, M., Rudnicky, A., Prasad, R., Aberdeen, J., Bratt, E. O., Garofolo, J., Hastie, H., Le, A., Pellom, B., Potamianos, A., Passonneau, R., Roukos, S., Sanders, G., Seneff, S., Stallard, D., 2002a. DARPA Communicator: Cross-System Results for The 2001 Evaluation. En: Proc. of the 7th International Conference on Spoken Language Processing (Interspeech'02-ICSLP). Vol. 1. Denver, USA, pg. 269–272.

Walker, M., Rudnicky, A., Prasad, R., Aberdeen, J., Bratt, E. O., Garofolo, J., Hastie, H., Le, A., Pellom, B., Potamianos, A., Passonneau, R., Roukos, S., Sanders, G., Seneff, S., Stallard, D., 2002b. DARPA Communicator Evaluation: Progress from 2000 to 2001. En: Proc. of the 7th International Conference on Spoken Language Processing (Interspeech'02-ICSLP). Vol. 1. Denver, USA, pg. 273–276.

Weizenbaum, J., 1966. ELIZA - A computer program for the study of natural language communication between man and machine. Communications of the ACM 9, 36–45.

Weng, F., Varges, S., Raghunathan, B., Ratiu, F., Pon-Barry, H., Lathrop, B., Zhang, Q., Scheideck, T., Bratt, H., Xu, K., Purver, M., Mishra, R., Raya, M., Peters, S., Meng, Y., Cavedon, L., Shriberg, L., 2006. CHAT: A

Conversational Helper for Automotive Tasks. En: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pg. 1061–1064.

Wilks, Y., 2006. Artificial companions as a new kind of interface to the future internet. Informe Técnico 13, Oxford Internet Institute.

Williams, J., Young, S., 2007. Partially Observable Markov Decision Processes for Spoken Dialog Systems. Computer Speech and Language 21(2), 393–422.

Wilting, J., Krahmer, E., Swerts, M., 2006. Real vs. acted emotional speech. En: Proc. of the 9th Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pg. 805–808.

Witten, I. H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Zeng, Z., Hu, Y., Fu, Y., Huang, T. S., Roisman, G. I., Wen, Z., 2006. Audio-visual emotion recognition in adult attachment interview. En: Proc. of the 8th International Conference on Multimodal interfaces. Banff, Alberta, Canada, pg. 828–831.

Zgank, A., Kaèiè, Z., Diehl, F., Vicsi, K., Szaszak, G., Juhar, J., Lihan, S., 2004. The COST278 MASPER initiative - Crosslingual speech recognition with large telephone databases. En: Proc. of the 4th International Conference on Language Resources and Evaluation (LREC 2004). Lisbon, Portugal, pg. 2107–2110.

Zong, Y., Dohi, H., Ishizuka, M., 2000. Multimodal presentation markup language mpml with emotion expression functions attached. En: Proc. of the 2nd International Symposium on Multimedia Software Engineering. pg. 359–365.

Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., Hetherington, L., 2000. JUPITER: A telephone-based conversational interface for weather information. En: IEEE Transactions on Speech and Audio Processing. Vol. 8. pg. 85–96.

Zue, V. W., Glass, J. R., 2000. Conversational interfaces: Advances and challenges. Proc. of the IEEE 88, 1166–1180.

La presente tesis doctoral describe el trabajo realizado en tres de las líneas más exigentes y prometedoras del área de los sistemas de diálogo oral: el reconocimiento de emociones no actuadas, la adaptación de reconocedores del habla entre idiomas y la evaluación de campo de sistemas. La investigación descrita en la Tesis constituye una aportación novedosa a lo que los expertos han definido como los mayores retos de investigación del área: la adaptatibilidad y portabilidad de los sistemas de diálogo oral.

En primer lugar, en cuanto al reconocimiento de emociones, se ha desarrollado una nueva aproximación eficiente que mejora considerablemente el reconocimiento de emociones no actuadas tanto automáticamente, como respecto a los niveles de acuerdo entre los anotadores humanos. Para ello, se propone el uso de diferentes fuentes de información contextuales.

En segundo lugar, durante una estancia de tres meses en la Technical University of Liberec (República Checa), se desarrolló e implementó un método eficiente en tiempo y esfuerzo para adaptar un reconocedor del habla a otros idiomas, que ha sido empleado con éxito para adaptar un reconocedor de voz checho a los idiomas eslovaco y español.

En tercer lugar, se han llevado a cabo diversos estudios estadísticos sobre la evaluación de campo de los sistemas de diálogo oral, proporcionando nuevas evidencias empíricas sobre las relaciones entre los diferentes criterios de evaluación. El estudio incluye tanto parámetros objetivos como subjetivos, prestando especial atención a la satisfacción del usuario y al éxito de la tarea.

Todas las propuestas de la Tesis se han evaluado con sistemas de diálogo reales, para lo cual se desarrolló el sistema de diálogo oral UAH.

On the Development of Adaptive and Portable Spoken Dialogue Systems:

Emotion Recognition, Language Adaptation and Field Evaluation

Zoraida Callejas Carrión

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Granada

ugr Universidad de Granada

On the development of Adaptive and Portable Spoken Dialogue Systems: Emotion Recognition, Language Adaptation and Field Evaluation

Zoraida Callejas Carrión

Dissertation submitted for the degree of
Doctor of Philosophy in Computer Science
with European Doctorate Mention

Supervised by:
Dr. Mr. Ramón López-Cózar Delgado

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Granada

Granada, April 2008

Acknowledgements

Some years ago I was introduced to the area of dialogue systems, and I was lucky it was by somebody who enjoys working, and enjoys working with them. Since then, in his enthusiasm I have had an splendid referent. If the role of a supervisor is to guide and advice, it can not be in any other way rather than encouraging to develop the own ideas freely. Thank you Ramón.

And to work freely it is necessary to go out of our laboratory, stop staring at our screen and find out what the others do, so that we can enrich our own contributions. Pasteur once said that science has got no nationality, but we scientists do. I would like to thank Jan Nouza and all the members of the Laboratoř počítačového zpracování řeči for demonstrating that this is not an obstacle, and that the only important thing to create a team is the willing to work towards shared objectives. Thank you for teaching me so many things, professionally and personally, and for inviting and receiving me in that spring I will never forget.

Also many thanks to Michael McTear. I remember when we discussed my first research results when I began my PhD studies, I started the venture of carrying out my Doctoral Thesis in a large extent because of his encouragement and interest. Thank you for offering me your valuable opinion and advice, which have made me improve, learn and progress.

I have also been lucky to share my day by day with wonderful people. I am very grateful to the colleagues in the department who have cheered me up, encouraged me in my work and congratulated me for my achievements, they have brighten up the way. Specially, I would like to thank my friends from Labs 1 and 2, including those who are no longer here, with whom I have lived the best moments and drank a lot of "coffee".

There are also some other coffees which, although sometimes "virtual", have rescued my from drowning. Those from my last resorts, from my friends who despite the distance have always demonstrated their affection. Thank you Bea, M^{a}Paz and Vane, without you, who understand me like nobody else, nothing would have been the same. In the distance my gratitude is also for Hynek, my guide in Hrabal's country, with your great sense of humor you have reduced the highest pitfalls: I will never forget your saviour bramborák.

My greatest thanks go to my conviction and security in the uncertainty moments: my family. To my parents and Clara for helping and believing in me constantly and unconditionally. Thank you for bearing the "side effects" of the thesis, the deadlines, the exhausting working days... This work is also yours. Thanks also to Carmen for sweetening my English classes since I was a child.

Finally, thank you David for your infinite patience, care and affection, for reminding me that it was worth and giving me wings to continue. Thank you for smoothing the way and holding my hand along it. Although you are the last I mention, *in my life* ...

Summary

This Thesis presents the work done on three of the most challenging topics in the area of spoken dialogue systems: recognition of non-acted emotions, cross-lingual adaptation of speech recognizers and field (not laboratory restricted) evaluation with both "objective" and "subjective" criteria. The research described constitutes a novel contribution to what the experts have established as the major research trends in the area of SDS: adaptiveness and portability of the systems.

Firstly, regarding emotion recognition, a detailed study is supplied on how to calculate and interpret reliability coefficients for the annotation of corpora of real emotions. A new efficient approach is proposed that considerably enhances inter-annotator agreement and machine emotion recognition by the use of several context information sources. On the one hand, human annotation is facilitated by achieving inter-annotator agreement values closer to the maximum attainable even with non-expert annotators. On the other hand, a machine-learned emotion recognition method is proposed which automatically extracts the contextual information at run time, obtaining results which provide a 40% improvement on the state-of-the-art approaches.

Secondly, the research on cross-lingual adaptation of speech recognizers was carried out during a three-month stay at the Technical University of Liberec (Czech Republic). An approach is presented to cost-efficiently (in terms of time and effort) adapt a speech recognizer to work in another language. The proposal has been used to adapt a Czech speech recognizer to a language which is acoustically very similar (Slovak) and another with a completely different origin (Spanish). It obtained a recognition accuracy around 70% for Spanish and 80% for Slovak in tasks demanding rich vocabularies (around 150,000 words).

Thirdly, several statistical studies were carried out on a field evaluation of a spoken dialogue system. New empirical evidence is provided on the relationships between evaluation criteria. The study includes both interaction parameters and quality judgments, paying special attention to user satisfaction and task success, studying the impact of the dialogue management initiatives employed, as well as the users expertise and collaboration during the interaction with the system.

All the methods proposed in the Thesis have been tested with real dialogue systems, for which the UAH spoken dialogue system was developed. The system has been available to the public on the phone from June 2005, since when all the interactions have been recorded. A year of user calls was semi-automatically annotated with de-facto standard evaluation criteria (e.g. word-error-rate). This corpus was extended with annotation of emotions by nine non-expert annotators, who tagged each utterance as "neutral", "angry", "doubtful" or "bored". Both the emotion recognition methods and the evaluation studies contained in this Thesis were tested with the UAH corpus. The proposed method for cross-lingual adaptation was evaluated using the MyVoice system, which was developed in the Technical University of Liberec. The translation of its commands for interaction in Spanish is another of the Thesis' contributions.

The empirical results obtained with the dialogue systems have been rigorously tested and their significance calculated using different statistical significance studies. The results of the research described have been published in several prestigious national and international conferences and journals. They have also formed part of oral presentations, posters and demonstrations internationally.

Contents

List of Tables

List of Figures

We must, with a view to the science which we are seeking, first recount the subjects that should be first discussed (...) for the subsequent free play of thought implies the solution of the previous difficulties, and it is not possible to untie a knot of which one does not know.

Aristotle, Metaphysics (Book III)

1 Introduction

Due to their constant improvement in performance and decreasing cost, computers have became an important part of our daily lives. We are surrounded by numerous electronic devices which provide information and functions which we are increasingly interested in accessing any time, anywhere and in our native language. Thus, new interfaces are needed to provide natural, intuitive and efficient ways of communication between humans and computers. Spoken language dialogue systems have emerged as a practical method for providing computers with intelligent communicative capabilities, as speech is the most natural and flexible means of communication among humans.

1.1 Spoken dialogue systems

A spoken dialogue system (SDS) is a software that accepts natural language as an input and produces natural language as an output engaging in a conversation with the user. To successfully manage the interaction with users, spoken dialogue systems usually carry out five main tasks: automatic speech recognition (ASR), natural language understanding (NLU), dialogue management (DM), natural language generation (NLG) and text-to-speech synthesis (TTS). These tasks are usually implemented in different modules. Figure 1.1 shows the typical modular architecture of an SDS.

1.1.1. Speech recognition

Speech recognition is the process of obtaining the text string corresponding to an acoustic input. It is a highly complex task, as there is a great deal of

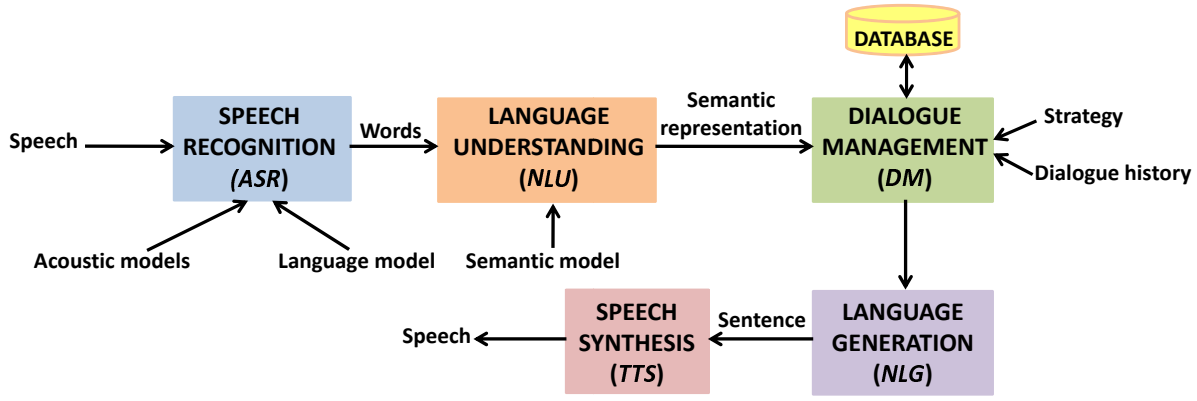


Figure 1.1. *Modular architecture of spoken dialogue systems*

variation in input characteristics, which can differ according to the linguistics of the utterance, the speaker, the interaction context and the transmission channel. Linguistic variability involves differences in phonetic, syntactic and semantic components that affect the voice signal. Inter-speaker variability refers to the wide differences between speakers regarding speaking style, voice, age, sex or nationality. Furthermore, even the same person does not always pronounce the same words in the same way, as people are affected by physical and psychological determinant factors that are highly variable and usually not predictable. This phenomenon is known as intra-speaker variability. Additionally, differences in the communication channels and/or devices also affect voice signals, due to effects derived from the transmission such as reverberation. Finally, the interaction environment variability is also very important, as the recognizer must be robust to differences in the background noise.

Different applications demand different complexity of the speech recognizer. Cole et al. (1997) identify eight parameters that allow optimal tailoring of the speech recognizer: speech mode, speech style, dependency, vocabulary, language model, perplexity, SNR and transductor. Regarding speech mode, speech recognizers can be classified into isolated-word or continuous-speech recognizers. The former recognize words separated by pauses, while the latter are able to recognize a natural discourse in which the speaker uses his normal speaking rate. Regarding speech style, a discourse can be read or spontaneous; the latter has peculiarities, such as hesitations and repetitions that make it more complex to recognize. Speech recognition can also be speaker-dependent or independent. In the first case the acoustic models

are trained with the voice of only one speaker, for whom optimal success rates are reached; whereas the second is prepared to recognize a wide range of speakers yielding acceptable success rates. Another important parameter is the accepted vocabulary, i.e. the number of words that the recognizer can distinguish. Applications with more than 5,000 accepted words are normally considered as large-vocabulary applications (Jurafsky and Martin, 2000). Finally, to deal with noise and channel variability, speech recognizers employ a noisy channel model, i.e. they treat the recognition problem as if it were necessary to recover a message which has been corrupted after going through a noisy channel. To do so, they use stochastic models which consider the possible original messages and calculate which of them is more likely to be the correct one.

1.1.2. Natural language processing

Once the SDS has recognized what the user uttered, it is necessary to understand what he said. Natural language processing is a method of obtaining the semantics of a text string and generally involves morphological, lexical, syntactical, semantic, discourse and pragmatical knowledge. In the first stage, lexical and morphological knowledge divide the words into their constituents by distinguishing between lexemes and morphemes: lexemes are parts of words that indicate their semantics and morphemes are the different infixes and suffixes that provide different word classes (e.g., $establishment = establish - ment$). Syntactic analysis yields the hierarchical structure of the sentences. However, in spoken language, phrases are frequently affected by difficulties associated with the so-called disfluency phenomena: filled pauses, repetitions, syntactic incompleteness and repairs (Gibbon et al., 2000). Semantic analysis extracts the meaning of a complex syntactic structure from the meaning of its constituent parts. In the pragmatic and discourse-processing stage, the sentences are interpreted in the context of the whole dialogue, the main complexity of this stage is the resolution of anaphora, and ambiguities derived from phenomena such as irony, sarcasm or double entendre.

There are currently two major approaches to tackling the problem of understanding: rule-based approaches and statistical models learned from data corpus. Rule-based approaches extract semantic information based on a syntactic-semantic analysis of the sentences, using grammars defined for the task, or by means of the detection of keywords with semantic meanings. In the case of statistical methods, the process is based on the definition of language units with semantic content and the learning of models from labelled samples. This kind of analysis uses a probabilistic model to identify concepts, markers and values of the cases, and uses them to represent the relationship between markers of cases and their values to semantically decode the user utterances (Minker, 1998).

1.1.3. Dialogue management

There is no universally agreed-upon definition of the tasks that a dialogue manager has to carry out. Traum and Larsson (2003) state that dialogue managing involves four main tasks: i) updating the dialogue context, ii) providing a context for interpretations, iii) coordinating other modules and iv) deciding the information to convey and when to do it. Thus, the dialogue manager has to deal with different sources of information such as the NLU results, database queries results, application domain knowledge, knowledge about the users and the previous dialogue history. Its complexity depends on the task and the dialogue flexibility and initiative. Bernsen et al. (1994) provide a taxonomy which shows that for small and simple tasks single-word dialogue can be convenient with either system or user initiative and limited system feedback. However, for large, well-structured tasks, there is a need for system-directed dialogues with appropriate system feedback, tracking of the dialogue history and simple user models. For larger ill-structured tasks, mixed initiative dialogues are necessary, with dynamic predictions, linguistic and dialogue act, dialogue history and advanced user modelling.

The simplest dialogue managing strategy is to model the dialogue as a finite-state machine in which the transitions between the system responses are determined by the user's actions. The users actions are his responses to the system, which are coded in recognition grammars. A significant extension consists of frame-based approaches, which have been developed to overcome the lack of flexibility of dialogue grammars. This is the approach used by

most current commercial systems. Unlike the finite-state approach, frame-based dialogue managers do not have a predefined dialogue path but use a frame structure comprised of one slot per piece of information that the system can gather from the user. In this approach, the system interprets the speech in order to acquire enough information to perform a specific action. Its advantage is that it can capture several data at once and the information can be provided in any order (more than one slot can be filled per dialogue turn and in any order).

For more complex domains, plan-based dialogue managing can be used. Its core idea is that humans communicate to achieve goals and during the interaction the mental state of the speakers may change. Thus, plan-based dialogue managers model dialogue as a cooperation between the user and the system to reach common goals, so that each utterance is not considered as a text string, but as a dialogue act in which the user communicates his intentions. From each user utterance the system refines a user model in which it tries to predict his intentions and objectives. This is done recursively to produce the system response to the user utterances until the task is accomplished.

This last approach is related to the so-called "information state" dialogue theory. The information state of a dialogue represents the information needed to uniquely distinguish it from all others. It comprises the accumulated user interventions and previous dialogue actions on which the next system response can be based. The information state is also sometimes known as the conversation store, discourse context or mental state. Following the information state theory, the main tasks of the dialogue manager are to update the information state based on the observed user actions, and based on them, to select the next system action. The Trindi project (TRINDIConsortium, 2001), proposed an architecture and toolkit for building dialogue managers based on an information state approach.

Additionally, when it is necessary to execute and monitor operations in a dynamically changing application domain, an agent-based approach can be employed. The modular agent-based approach to dialogue management makes it possible to combine the benefits of different dialogue control models, such as finite-state based dialogue control and frame-based dialogue managing (Chu et al., 2005). Similarly, it can benefit from alternative dialogue management strategies, such as the system-initiative approach and the

mixed-initiative approach (Walker et al., 1998a). It also makes it possible to combine rule-based and machine learning approaches (Turunen et al., 2004).

The application of machine-learning approaches to dialogue management strategy design is a rapidly growing research area. Machine-learning approaches to dialogue management attempt to learn optimal strategies from corpora of real human-computer dialogue data using automated "trial-and-error" methods instead of relying on empirical design principles (Griol, 2007). The Markov-Decision-Process (MDP) model serves in most of these approaches as a formal representation of human-machine dialogue and provides the basis for formulating strategy learning problems (Williams and Young, 2007; Cuayáhuitl et al., 2006; Lemon et al., 2006).

1.1.4. Natural language generation

Natural language generation is the process of obtaining texts in natural language from a non-linguistic representation. It is usually carried out in 5 steps: content organization, content distribution in sentences, lexicalization, generation of referential expressions and linguistic realization. It is important to obtain legible messages, optimizing the text using referring expressions and linking words and adapting the vocabulary and the complexity of the syntactic structures to the user's linguistic expertise.

The simplest approach consists of using predefined text messages (e.g. error messages and warnings). Although intuitive, this approach completely lacks from flexibility. The next level of sophistication is template-based generation, in which the same message structure is produced with slight alterations. The template approach is used mainly for multi-sentence generation, particularly in applications whose texts are fairly regular in structure, such as business reports.

Phrase-based systems employ what can be considered as generalized templates at the sentence level (in which case the phrases resemble phrase structure grammar rules), or at the discourse level (in which case they are often called text plans). In such systems, a pattern is first selected to match the top level of the input, and then each part of the pattern is expanded into a more specific one that matches some portion of the input. The cascading process stops when every pattern has been replaced by one or more words.

Finally, feature-based systems represent the maximum level of generalization and flexibility. In feature-based systems, each possible minimal alternative of expression is represented by a single feature; for example, whether the sentence is either positive or negative, if it is a question or an imperative or a statement, or its tense. To arrange the features it is necessary to employ linguistic knowledge. Another alternative is to use corpus-based natural language generation (Oh and Rudnicky, 2000), which stochastically generates system utterances.

1.1.5. Text to speech synthesis

Text-to-speech synthesizers transform a text into an acoustic signal. A text-to-speech system is composed of two parts: a front-end and a back-end. The front-end carries out two major tasks. Firstly, it converts raw text containing symbols such as numbers and abbreviations into their equivalent words. This process is often called text normalization, pre-processing, or tokenization. Secondly, it assigns a phonetic transcriptions to each word, and divides and marks the text into prosodic units, i.e. phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme or grapheme-to-phoneme conversion. The output of the front-end is the symbolic representation constituted by the phonetic transcriptions and prosody information.

The back-end (often referred to as the synthesizer) converts the symbolic linguistic representation into sound. On the one hand, speech synthesis can be based on human speech production. This is the case of parametric synthesis which simulates the physiological parameters of the vocal tract, and formant-based synthesis, which models the vibration of vocal chords. In this technique, parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech. Another approach based on physiological models is articulatory synthesis, which refers to computational techniques for synthesizing speech based on models of the human vocal tract and the articulation processes.

On the other hand, concatenative synthesis employs pre-recorded units of human voice. Concatenative synthesis is based on stringing together segments of recorded speech. It generally produces the most natural-sounding synthesized speech; however, differences between natural variations in speech

and the nature of the automated techniques for segmenting the waveforms sometimes result in audible glitches in the output. The quality of the synthesized speech depends on the size of the synthesis unit employed. Unit selection synthesis uses large databases of recorded speech. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, syllables, morphemes, words, phrases, and sentences. Unit selection provides the greatest naturalness, because it applies only a small amount of digital signal processing to the recorded speech. There is a balance between intelligibility and naturalness of the voice output or the automatization of the synthesis procedure. For example, synthesis based on whole words is more intelligible than the phone-based but for each new word it is necessary to obtain a new recording, whereas the phones allow building any new word. In one extreme, domain-specific synthesis concatenates pre-recorded words and phrases to create complete utterances. It is used in applications in which the variety of texts the system will produce is limited to a particular domain, like transit schedule announcements or weather reports. At the other extreme, diphone synthesis uses a minimal speech database containing all the diphones (sound-to-sound transitions) occurring in a language. The number of diphones depends on the phonotactics of the language: for example, Spanish has about 800 diphones and German about 2,500. In diphone synthesis, only one example of each diphone is contained in the speech database.

Finally, HMM-based synthesis is a method in which the frequency spectrum (vocal tract), fundamental frequency (vocal source), and duration (prosody) of speech are modelled simultaneously by HMMs. Speech waveforms are generated from HMMs themselves, based on the maximum likelihood criterion.

1.2 Applications of SDSs

The complexity of the interaction between the user and the dialogue system can vary and some of the previously described components might not be used. For example, for a simple menu, semantic analysis is not necessary. However, for a conversational companion all the modules must be used in order to interpret the user input, take justified decisions on what the system will respond, and finally tailor the answer to user needs and expectations.

There is a large variety of applications in which spoken dialogue systems can be used. One of the most wide-spread is information retrieval. Some sample applications are:

- DARPA Communicator - Intelligent conversational interfaces to distributed flight and booking information (DARPA, 1992, 1994).
- Voyager - Tourist and travel information for the Greater Boston area (Glass et al., 1995).
- ARISE - Automatic Railway Information Systems for Europe in several languages (den Os et al., 1999).
- AUGUST - Swedish spoken dialogue system using an animated agent to provide information about Stockholm (J. Gustafson and Lundeberg, 1999).
- Adapt - Multimodal spoken dialogue system for browsing apartments on the Stockholm real estate market (Gustafson et al., 2000).
- Jupiter - Weather forecast over the phone (Zue et al., 2000).
- Mercury - Flight reservation system and weather forecast (Seneff and Polifroni, 2000).
- CTT-Bank - Speech-controlled telephone banking system (Melin et al., 2001).
- SmartKom - German multimodal dialogue system with several application domains such as cinema booking (Alexandersson and Becker, 2001).
- Let's Go - A spoken dialogue system for the non-native and elderly in the domain of bus information around Pittsburgh (Raux et al., 2005).
- Amities project - Automated Multilingual Interaction with Banking Information and Services (Hardy et al., 2006).
- DisCoH - Spoken Dialogue System for Conference Help (Andeani et al., 2006).

- HIGGINS - Pedestrian city navigation and guidance (Skantze et al., 2006).
- TALK TownInfo - Multimodal dialogue system using reinforcement learning for tourist information scenarios (Lemon et al., 2006).
- Conquest - Spoken dialogue system that provides schedule information during conferences (Bohus et al., 2007).

Spoken dialogue systems have also been used for education and training, particularly in improving phonetic and linguistic skills:

- LARRI - Multimodal spoken dialogue system which provides assistance and guidance to F18 aircraft personnel during maintenance tasks (Bohus and Rudnicky, 2002).
- ITSPOKE - Tutoring spoken dialogue system which engages the students in a spoken dialogue to provide feedback and correct misconceptions (Litman and Silliman, 2004).
- Radiobot-CFF - Spoken dialogue system which can engage in Call For Fire (CFF) radio dialogues to help train soldiers in proper procedures for requesting artillery fire missions (Roque et al., 2006b).
- VOCALIZA - Dialogue application for computer-aided speech therapy in the Spanish language, which helps in the daily work of speech therapists who teach linguistic skills to Spanish speakers with different language pathologies (Vaquero et al., 2006).
- LISTEN - "Literacy Innovation that Speech Technology ENables" is an automated Reading Tutor that displays stories on a computer screen, and listens to children read aloud (Mostow, 2008).

In some cases, spoken interaction can be the only way to access information, as, for example when the screen is too small to display information (e.g. hand-held devices) or when the eyes of the user are busy in other tasks (e.g. driving):

- MUST - Multimodal, multilingual information services for small mobile Terminals (Boves and Os, 2002).

- VICO - Virtual Intelligence CO-driver enabling natural interaction between humans and digital devices and services in the car (Mattasoni et al., 2002).
- Athosmail - A multilingual adaptive spoken dialogue system for e mail message reading on a mobile phone (Jokinen et al., 2004).
- CHAT - Conversational helper for automotive tasks such as query song databases and operate an MP3 player (Weng et al., 2006).
- DICO - A multimodal dialogue system that lets the driver control devices and access Internet services using natural speech (Villing and Larsson, 2006).

Spoken interaction is also useful for remote control of devices and robots, specially in smart environments:

- WITAS - Dialogue interface for multimodal conversations with WITAS, a robotic helicopter (Lemon et al., 2001).
- ODISEA - A spoken dialogue system that allows interaction between users and intelligent environments. The dialogue components are automatically created and permit context-based spoken interaction between the environment and users (Montoro et al., 2004, 2006).
- SENECA - Speech-based user interface in a wide range of entertainment, navigation and communication applications in mobile environments by means of human-machine dialogues (Minker et al., 2004a).
- Clarissa - Fully voice-operated procedure browser, allowing astronauts to make more efficient use of their hands and eyes so as to give full attention to the task while they navigate through the procedure using spoken commands (Rayner et al., 2005).
- MIMUS - Multimodal dialogue system for the control of a smart home. It relies on a flexible architecture that allows for the integration of multiple input and output modalities (Pérez et al., 2006).
- STanford AI Robot (STAIR) - Robotic assistant, capable of conversation, for home and office (Krsmanovic et al., 2006).

- Cogniron - The cognitive robot companion (Menezes et al., 2007).

Finally, one of the most demanding applications for fully natural and understandable dialogues are virtual agents and companions:

- Collagen - Building conversational assistants and collaborative agents (Rich and Sidner, 1998).
- AVATALK - Natural, interactive dialogues with responsive virtual humans (Hubal et al., 2000).
- COMIC - Bathroom design using speech and gesture input/output, in collaboration with an avatar with facial emotions (Catizone et al., 2003).
- NICE - Embodies historical and literary characters capable of natural fun and experientially rich communication with children and adolescents (Corradini et al., 2004).

1.3 Evolution of SDSs

Human beings have always wanted to be able to communicate with artificial companions. There are many examples in cinema and literature. Some of the most ancient examples can be found in Greek and Roman mythology in which heroes could communicate with statues of goddesses or warriors. The first serious attempts at building talking systems were initiated in the 18th and 19th centuries, when the first automata were built to imitate human behaviour. The first of these were clockwork machines in which the masters applied all their skill to building animals or dolls that could produce sounds. In 1770, Baron Von Kempelen developed the first automaton that produced whole words and short phrases, which was subsequently improved by Josef Faber, who built the Euphonia machine in 1857. Euphonia imitated the mechanism of human speech by the use of a bellows which pumped air through a series of plates and chambers that could modulate the sounds by employing a 16-keys keyboard similar to a piano. The machine could speak any word in several European languages. These first machines were mechanical, and it was not until the end of the 19th century when scientists concluded that speech could be produced electrically.

At the beginning of the 20th century, J.Q. Stewart (Stewart, 1922) built a machine that could generate vocalic sounds electrically. During the 30s, the first electric systems were built that could produce any type of sound. The first was the VOCODER, a speech analyser and synthesizer developed by Bell Laboratories that could be operated by a keyboard. A skilled human operator could select either a periodic source for sonorant sounds or a noise source for fricative sounds that could be altered by controlling a filter bank. At the same time the first systems appeared with very basic natural language processing capabilities for machine translation applications.

During the 40s, the first computers were developed and some prominent scientists like Allan Turing pointed out their potential for applications demanding "intelligence". To measure a machine's capability to demonstrate intelligence, Turing (1950) proposed the so-called "Turing test" in which a human judge engaged in a natural language conversation with the machine. If the judge was not able to reliably tell whether he had talked with a man or a machine, the machine passed the test. This was the starting point that fostered the research initiatives that in the 60s yielded the first conversational agents. For example Weizenbaum's ELIZA (Weizenbaum, 1966), which was based on keyword spotting and predefined templates. The templates allowed the user input to be transformed into system answers. For example, when the user wrote a sentence such as "I am X", ELIZA would reply "How long have you been X?" independently of the meaning of 'X'. Thus, although their behaviour was perceived as human by some naïve users and they might pass the Turing test, in practice the first conversational systems such as ELIZA did not semantically interpret the users' input. To address this challenge, the research area of computational linguistics appeared in the 70s, grounded on the theoretical work developed in the 50s by Chomsky, Montague and Wood. The first rule-based speech synthesizers appeared about the same time. In the 70s the first continuous speech recognizers also appeared. These were based on decades of research on discrete speech in which verbal stimuli were punctuated by long pauses.

Benefiting from the continual improvements in the areas of speech recognition, natural language processing and speech synthesis, the first research initiatives related to spoken dialogue systems appeared in the 80s. To some extent the origin of this research area is linked to two seminal projects: the DARPA Spoken Language Systems programme in the USA and the Esprit SUNDIAL in Europe. On one hand, the main objective of the DARPA project was the study and development of technologies related to ASR and NLU in the domain of flight reservation by phone, to which they gae the name Air Travel Information Services (ATIS) (DARPA, 1992) (DARPA, 1994). The ATIS dialogue corpus, which is still employed by SDS developers and researchers, was followed by other projects, such as those carried out by AT&T, for example the AMICA project (Pieraccini et al., 1997), in which different stochastic models were applied to a mixed-initiative SDS. ATIS was also the starting point for research in MIT and CMU, where some of the most important systems in academia have been created. The SUNDIAL project was concerned with flight and train timetables in four different European languages. The research carried out in SUNDIAL yielded numerous projects funded by the European Community and mainly concerned with dialogue modelling, such as VERMOBIL (Bos et al., 1999), DISC (Bernsen and Dybkjaer, 1997) and ARISE (den Os et al., 1999). In the ARISE project six different systems were developed simultaneously: two Italian prototypes based in the technologies developed in CSELT (Castagneri et al., 1998) (Baggia et al., 2000), a French prototype developed by LIMSI (Lamel et al., 2000b) and two prototypes in Dutch and French based on the Philips technology. The DARPA ATIS project has been considered by some authors (Bangalore et al., 2006) to have been included in an earlier generation of SDSs than the SUNDIAL, as it was restricted to a closed application domain.

Among the most important research programmes of the 90s with multi-domain capabilities, the DARPA Communicator stands out. This government-funded project aimed at the development of cutting-edge speech technologies, which could employ as an input not only speech but also other modalities. The systems developed in this programme by both US and European partners were able to engage in complex interactions with the users in multiple domains, in which either the user or the system could begin the conversation, change topic or interrupt the other. For example, the CMU researchers developed the Carnegie Mellon Communicator system (Rudnicky

et al., 1999), which provided information on complex itineraries which included booking multiple flights, hotels and car rentals. The system architecture was based on specialized agents which developed modules that worked independently, encapsulating task-dependent information.

The most important research guidelines for the 90s were related to improving the success rates of the different components of the dialogue systems. Firstly, regarding speech recognition, their main concern was robustness. Authors argued about the sudden performance degradation in the systems due to minor changes, such as changing the microphones or the telecommunications channels used, and stated that the state-of-the-art technology used at that time was not capable of providing acceptable solutions. In order to overcome these problems, research was mainly focused on fundamental issues such as robustness (Cole et al., 1995), studying how to model the spectral characteristics (Holmes and Huckvale, 1994; Ostendorf et al., Sep 1996), enhancing coarticulation models (Sun, 1997; Kirchhoff and Bilmes, 1999), modelling speech rate (Pfau and Ruske, 1998; Morgan et al., 1997) and providing speaker-independent recognition methods to cope with speaker differences. Robust speech recognition was acknowledged to be one of the most important aims in the context of the DARPA ATIS domain Stern et al. (1992). In Europe, the COST Action 249, which took place that took place from 1994 to 2000[1], with a research team from 20 European countries, was concerned with continuous ASR over the telephone, and covered all the previous topics, including selection of acoustic models, phonetic classification methods and adaptation to the characteristics of the telephone links.

In the late 90s, systems had to be enhanced with the ability to cope with differences between landline and cellular phones, as the latter were then becoming popular. The new systems required the ability to handle narrow channel bandwidths and low signal-to-noise ratios. Additionally, the adoption of cellular phones involved dealing with an increasingly richer variety of environments from which the users contacted with the systems and thus had to be robust to handle communication in very noisy backgrounds (Kacic, 1999). For example, one of the main topics studied for the development of the TREC-8 SDR system in LIMSI was noise compensation (Gauvain, 1999).

The main research guidelines in the field of NLU during the 90s were related to work with richer vocabularies, thus moving from isolated word

[1]http://www.elis.ugent.be/cost249/

recognition applications to spontaneous speech. In order to do so, the authors conducted research on how to deal with out-of-vocabulary words, establishing successful and efficient vocabulary learning as a desirable future achievement. In this case, proposals were also related to studying highly specific methods and algorithms, such as enhancing linguistic analyses or trying to avoid them using more sophisticated word-spotting techniques (Zue and Glass, 2000).

In the 90s, great efforts were made in dialogue management towards the use of less restricted dialogues in which users could take the initiative in the communication. Thus, authors claimed for barge-in to be considered not only from the speech recognition point of view but also from the interaction perspective (Zue and Glass, 2000). Bangalore et al. (2006) describe three generations of spoken dialogue systems regarding, among other characteristics, the dialogue management initiatives and natural language understanding capabilities. Firstly, they describe a first generation in which SDSs used system-directed initiatives and the semantics were directly associated with the detection of keywords. The second generation was comprised of mixed-initiative dialogue systems in which the natural language understanding was carried out using frames. This allowed the users to talk naturally about a single task. Third generation SDSs additionally supported multiple tasks or domains simultaneously and could be enhanced with multi-modal and multi-media capabilities.

One of the major trends during the 90s was related to the definition of standard languages for the development of SDSs (Zue and Glass, 2000). For example, at the end of 1999 the W3C Voice Browser working group presented the first requirement studies for web browsers which settled the basis for the requirements of markup languages for spoken dialogues[2]. The modularization of the systems to obtain more portable and re-usable components was mainly done by obtaining sets of frequent sub-dialogues. It was not until the late 90s when the first architectures for developing plug-and-play components appeared. For example, in 1998 the Galaxy architecture appeared, one of the pioneer works in fostering development of completely independent components for SDS (Seneff et al., 1998).

[2]http://www.w3.org/TR/voice-dialog-reqs/

There were also a generally positive attitude towards the adoption of stochastic approaches to obtain unsupervised methods to augment ASR and NLU capabilities (Glass, 1999). Thus, there was an increasing need for shared linguistic resources and data collections with which to train their algorithms. During the 90s, the first corpora development and system evaluation frameworks appeared yielding big shared resources such as WordNet.

Throughout the history of spoken dialogue systems, some experts have dared to envision what the future research guidelines in the area would be (see Figure 1.2). These objectives have gradually changed towards ever more complex goals. As reflected in the previously mentioned research results, during the 90s the major trends were towards making all system components (ASR, NLU, DM, TTS) more robust (Cole et al., 1995; Kacic, 1999; Zue and Glass, 2000; Mangold, 2001). Since 2003, experts have proposed higher level objectives, such as providing the system with advanced reasoning, problem solving capabilities, adaptiveness, proactiveness, affective intelligence, multimodality and multilinguality (Dale, 2003; Jokinen, 2003; Gao et al., 2005; Haas et al., 2005; Minker et al., 2006b,a). As can be observed, these new objectives refer to the system as a whole, and represent major trends that in practice are achieved through joint work in different areas and components of the dialogue system. For example, for a system to be multilingual, it has to be able to recognize the users' utterances in different languages, interpret their semantics using the corresponding linguistic knowledge, and also generate natural language and synthesize speech in all of them. Thus, in contrast to what happened in the 90s, when each area had different objectives (ASR, NLU, DM, NLG and TTS), current research trends are characterized by large-scale objectives which are shared out between the different researchers in different areas.

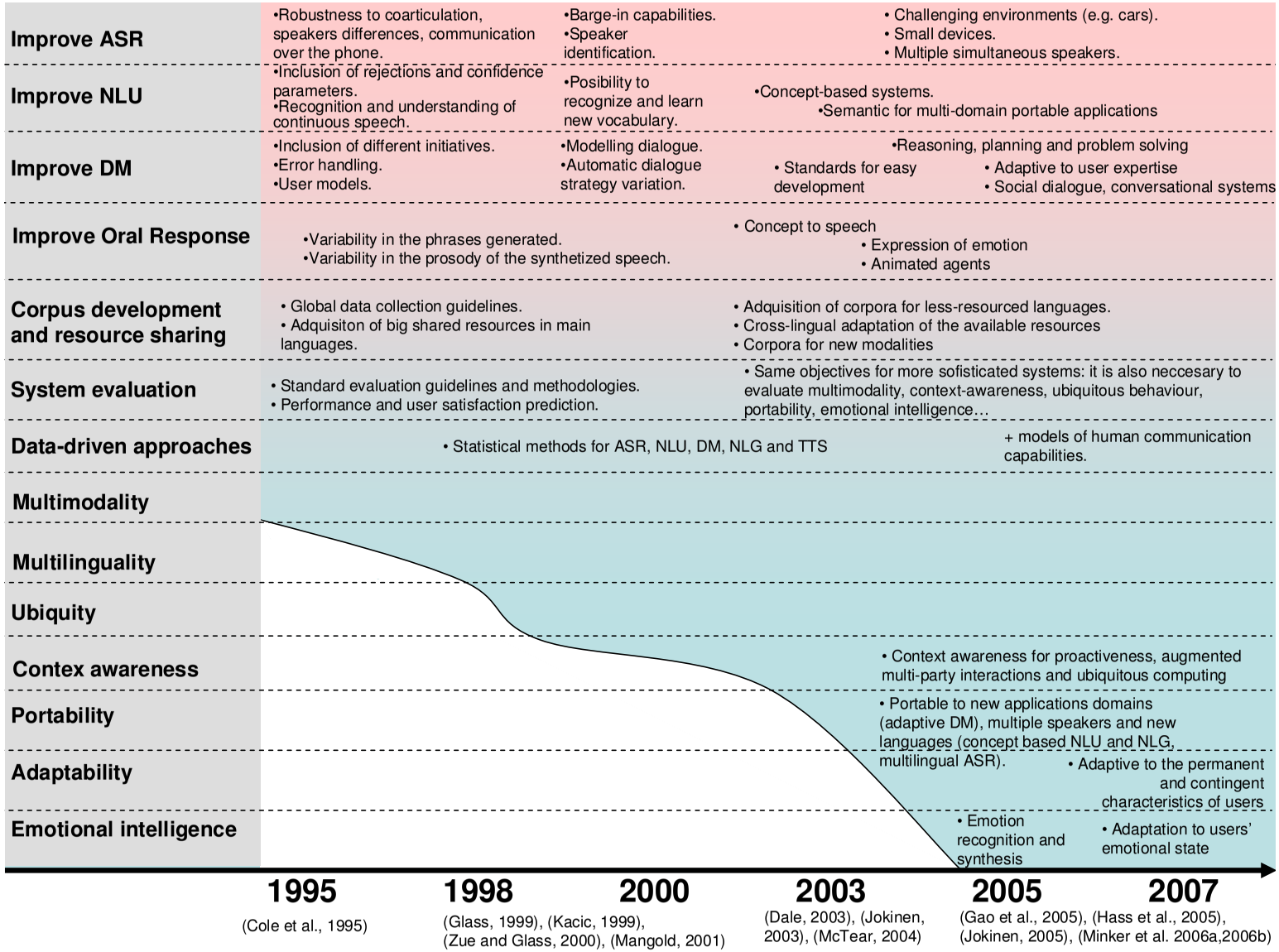


Figure 1.2. *Evolution of the envisioned research lines in the area of spoken dialogue systems*

Specialists have recently envisioned future dialogue systems as being intelligent, adaptive, proactive, portable and multimodal. All these concepts are not mutually exclusive, as for example the system's intelligence can also be involved in the degree to which it can adapt to new situations, and this adaptiveness can result in better portability for use in different environments.

Proactiveness is necessary for computers to stop being considered a tool and becoming real conversational partners. Proactive systems have the capability of engaging in a conversation with the user even when he has not explicitly requested the system's intervention. This is a key aspect in the development of ubiquitous computing architectures in which the system is embedded in the user's environment, and thus the user is not aware that he is interacting with a computer, but rather he perceives he is interacting with the environment. In such situations, proactiveness enables the system to passively observe the dialogue between human users and capture the relevant conversational context, which the system processes to compute when the conversational situation requires it to take the initiative and get meaningfully involved in the communication. To achieve this goal, it is necessary to provide the systems with problem-solving capabilities and context-awareness. For example, in Schneider (2004) presents a proactive shopping assistant integrated in supermarket trolleys. The system observes the shopper's actions and tries to infer the user's goals. With this information, it proactivelly offers adapted support tailored to the current context, as for example displaying information about products when the user holds them for a very long time or comparing different products when the user is deciding between two items. Other proactive systems are described in (Baudoin et al., 2005; Kwon et al., 2005).

The interest in developing systems capable of maintaining a conversation as natural and rich as a human conversation, gave rise to research on multimodal interfaces. Unlike traditional keyboard and mouse interfaces or unimodal speech dialogue systems, with multimodal interfaces there is flexibility in the input and output modes such as speech gesture or facial expressions. As multimodality permits users to employ different input modalities as well as to obtain responses through different vias, it is specially important for users with special needs, for which the traditional interfaces are not suitable. The first multimodal dialogue systems appeared in the mid-nineties, basically combining speech with pointing maps (Cheyer and Julia, 1995; Oviatt

et al., 1997) and speech with pen input (Waibel et al., 1997). Eye movements and gaze were one of the first modalities being studied, in order to identify which objects the user was referring to in his conversation (Sarukkai and Hunter, 1997). Most of these systems focus on using multimodality in the input or output, but some recent large projects have been directed towards the development of full multimodality. This is the case of the SMARTKOM project, which provides what is called "full symmetric multimodality" in a mixed-initiative dialogue system (Wahlster, 2006). They define symmetry as the capability of the system not only to understand and represent the user's modal input but also to create multimodal output. The main contribution of the Smartkom project is that it not only deals with modality integration or synchronization but also covers the dialogue phenomena that are associated with multimodality, such as mutual disambiguation, multimodal deixis and cross-modal reference resolution and generation, multimodal anaphora and ellipsis resolution and generation and multimodal turn-taking.

Adaptivity may also refer to other aspects in speech applications. In speech-based human-computer interaction users have diverse ways of communication. Novice users and experienced users may want the interface to behave completely differently, for example to have system-initiative instead of mixed-initiative. An example of the benefits of adaptivity in the interaction level can be found in (Litman and Pan, 2002). Multilingual applications are another example of adaptive applications. Multilingual recognizers are capable of recognizing simultaneously several languages by sharing acoustic and/or language models. Multilingual acoustic models consist of either a collection of language-dependent acoustic models for each language, or a combination of language-independent acoustic models (Schultz and Kirchhoff, 2006). The development of a speech recognizer is a very arduous and time demanding task. A large amount of data spoken by hundreds of subjects must be recorded and carefully annotated to get a representative set suitable for training the acoustic models. To overcome this problem there has been increasing interest in developing rapid prototyping approaches. Cross-lingual approaches have arisen in order to share language-recognition resources. This is particularly interesting for the development of speech recognizers for the less-resourced languages, which is a very important research area nowadays, and many of the main conferences in speech technologies explicitly devote sessions to studying this topic, as for example the Interspeech 2007 Spe-

cial Session on Speech and language technology for less-resourced languages (2007) and the LREC 2008 Workshop on Collaboration: interoperability between people in the creation of language resources for less-resourced languages (2008).

As stated by Jokinen (2003), there are different levels in which the system can adapt to the user. The simplest one is through personal profiles in which the users have static choices to customize the interaction (e.g. whether they want a male or female system's voice), which can be further improved by classifying users into preferences' groups. Systems can also adapt to the users' environment, for example ambience intelligence applications such as the ubiquitous proactive systems described. A more sophisticated approach is to adapt to the user's knowledge and expertise. The main research topics are the adaptation of systems to different user expertise levels (Haseel and Hagen, 2005), user knowledge (Forbes-Riley and Litman, 2004b) and user special needs. This last topic is receiving a lot of attention in terms of how to make systems usable by handicapped people (Heim et al., 2007), children (Batliner et al., 2004) and the elderly (Langner and Black, 2005) and also to adapt to permanent features like users' age, proficiency in the interaction language (Raux et al., 2003) or the user's expertise in using the system (Haseel and Hagen, 2005). Despite their complexity, these characteristics are to some extent rather static, Jokinen identifies another degree of adaptation in which the system not only adapts to the explicit message conveyed during the interaction, but also to the user's intentions and state. Following this guideline, affective computing studies focus on how to recognize and adapt to the user's emotional state during the conversation with the system.

There is an increasing interest in the development of spoken dialogue systems that dynamically adapt their conversational behaviours to the users' affective state. Martinovski and Traum (2003) demonstrated by means of user dialogues with a training system and a telephone-based information system that many breakdowns in man-machine communication could be avoided if the machine was able to recognize the emotional state of the user and responded to it more sensitively. Earlier experiments by Prendinger et al. (2003) showed that an empathetic computer agent can indeed contribute to a more positive perception of the interaction. Polzin and Waibel (2000) present a first approach to adjusting dialogue behaviour to the user's emotional state. For instance, they suggest that more explicit feedback should

be given if the user is frustrated. Nevertheless, their approach relies on few selection rules and is not based on a general framework for affective dialogue. Walker et al. (1997) examined how social factors, such as status, influence the semantic content, the syntactic form and the acoustic realization of conversations. Emotional intelligence includes the ability to recognize the user's emotional state as well as the ability to act on it appropriately. At present, there are a number of projects whose main objective is to endow dialogue systems with emotional intelligence. Some of the latest are MEGA (Camurri et al., 2004), NECA (Gebhard et al., 2004), VICTEC (Hall et al., 2005), NICE (Corradini et al., 2005), HUMAINE (Cowie and Schröder, 2005) and COMPANIONS (Wilks, 2006), to mention just a few.

Portability is currently addressed from very different perspectives, the three main ones being domain, language and technological independence. Ideally, systems should be able to work over different application domains, or at least be easily adaptable between them. Current studies on domain independence centre on how to merge lexical, syntactic and semantic structures from different contexts (Chambers and Allen, 2004) and how to develop dialogue managers that deal with different domains (Mourão et al., 2004; Nguyen and Wobcke, 2006). Regarding language independence, multilingual systems (Schultz and Kirchhoff, 2006) are those which can work with several languages and are thus portable in two main senses: firstly, users can input information in different languages, and secondly they can also receive the response in different languages. This is specially useful in speech-to-speech systems, which can serve as real time interpreters, so that as the first speaker talks on the telephone, the other receives the information translated into another language.

Finally, technological independence deals with the possibility of using dialogue systems with different hardware configurations. Computer processing power will continue to increase, with lower costs for both processor and memory components. The systems that support even the most sophisticated speech applications will move from centralized architectures to distributed configurations and thus must be able to work with different underlying technologies. To achieve this objective, different standard architectures have been studied. The main alternative, as gathered from the efforts of both universities and business companies, is that proposed by the MMI Working Group

in W3C[3], whose goal is to provide a general and flexible framework ensuring interoperability among modality-specific components from different vendors and underlying technologies.

1.4 Objectives of the Thesis

The main goal of the Thesis was to contribute to fostering the adaptiveness and portability of SDSs, which are the main current trends established by the experts in the area. The work presented in this Thesis consists of the development of models and methods for handling several of the issues described above, and experimentally test their performance with real dialogue systems. Of course, all aspects could not be included in the scope of the Thesis and it was decided to study three of the most challenging topics: emotion recognition, cross-lingual adaptation and field evaluation.

- Emotion recognition. The main objective was to find decisive factors in the recognition of emotions that influence both human and machine-learned recognition. Most state-of-the-art approaches are based on the use of simulated emotions in order to have strict control over the corpora used to train the automatic procedures. The aim of the Thesis was to use natural user interactions with a real dialogue system. This is a particularly challenging objective as in this case emotions are produced more subtly and the proportion of emotional utterances is very low compared to neutral cases. The Thesis centres on studying the recognition of negative emotions that can make the interaction with SDSs fail.

- Cross-lingual adaptation. One of the main claims of the research community has been the need for common resources and for ways of making the most of those available. This is vital for research centred on minority languages or dialects. The objective was to develop a technique that would allow fast adaptation of a speech recognizer to work in another language without the need of building new acoustic models. To maximize portability the goal was to make it employable not only between languages with similar origins, but also with very different languages.

[3]http://www.w3.org/TR/mmi-arch/

- Field evaluation. To be able to build systems that adapt to the users' needs and expectations it is important to take into account their opinions about previous interactions with the systems. In the literature, most evaluations are carried out with laboratory studies and consider interaction parameters and quality judgements separately. The main objective was to discover significant relationships between both interaction parameters and quality judgements in a field study on non-restricted user interactions of the users with a real system.

1.5 Structure of the Thesis

After this concise overview of the main research challenges for the creation of adaptive and portable systems, the contributions of the Thesis are presented on chapters 2 to 5. Each of the chapters is structured as follows: they start with an introduction which explains the objective of the research carried out; then a detailed state-of-the art of the topic is presented and the Thesis contributions are compared with prior work in the dialogue literature. The Thesis relies to a large extent on an empirical approach in which all the proposals are evaluated with real dialogue systems. In each of the chapters there is a section devoted to explaining the experimental set-up, followed by others that present the experimental results and conclusions extracted from them.

Chapter 2 describes the UAH spoken dialogue system. The system was carefully developed using dynamic dialogue management and different dialogue initiatives and confirmation strategies. The objective was to study the effect of each of them on the user's opinions about interactions with the system. In this chapter, the functionality and structure of the UAH system is discussed, giving a detailed description of each of its modules and highlighting its innovative functionalities. The system is accessible to the general public on the telephone and the users' calls were recorded and later annotated. The resulting corpus was used for the research on the field evaluation of SDSs described in chapter 5 where the annotation procedure is roughly described. The UAH corpus was also employed for the emotion recognition research in chapter 3. Thus, UAH was developed using innovative approaches and was also a testbed for the different proposals, i.e. the approaches proposed in the Thesis have been successfully employed in a real SDS.

Chapter 3 presents the research done on emotion recognition and includes a detailed study on the impact of considering context information for the annotation of emotions. The inclusion of the history of user-system interaction and the neutral speaking style of users is proposed. From the human annotation perspective, special attention is paid to how to reliably compute inter-annotator agreement. From the machine-learning perspective, a new method to automatically include both sources of information has been developed, making use of novel techniques for acoustic normalization and dialogue context annotation. The proposals were tested with the corpus extracted from the interactions of approximately 60 users with the UAH system.

Chapter 4 describes the research on cross-lingual adaptation, introducing a methodology that proved to be successful and cost-efficient for porting an existing speech recognition system to other languages is proposed, which was tested with the MyVoice system[4]. The methodology was tested with three different languages, Czech, Spanish and Slovak, to find out whether it was possible to use the proposed method not only with languages of similar origin (in this case Czech and Slovak), but also with languages that belong to very different branches of the Indo-European language family (such as Czech and Spanish). The proposed approach consists of four steps: 1) creating an initial mapping between the phonemes in the original language and the target language, 2) creating a lexicon and automatically mapping the words' pronunciation to the original phoneme set, 3) fine-tuning the models of the phonemes that are unique to each particular language, and 4) collecting data in the new target language and carrying out speaker adaptation. The research described in this chapter was done in collaboration with researchers from the Technical University of Liberec (Czech Republic), during a three-month stay in the Laboratory of Computer Speech Processing[5] under the supervision of Prof. Jan Nouza.

[4]MyVoice was developed in the Technical University of Liberec (Czech Republic)

[5]http://itakura.kes.tul.cz/kes/indexe.html

Chapter 5 shows the work done on the field evaluation of the UAH system. Evaluation of spoken dialogue systems has been traditionally carried out in terms of instrumentally or expert-derived measures (usually called "objective" evaluation), and quality judgments of users who have previously interacted with the system (also called "subjective" evaluation). Different research efforts were made to extract relationships between these evaluation criteria. In this chapter empirical results obtained from statistical studies are reported. These studies were carried out on interactions of real users with a real SDS, something which is rarely found in the literature.

The Thesis ends with chapter 6, which presents the overall conclusions, gives a detailed overview of the main contributions, and summarizes the main guidelines for future work. Finally, the Appendix presents the publications referring to the research described in the Thesis.

Puisque c'est ell que j'ai écoutée se plaindre,
our se vanter, ou même quelquefois se taire.
Puisque c'est ma rose

Antoine de Saint-Exupery, Le Petit Prince

2 The UAH spoken dialogue system

2.1 Introduction

In this chapter, the Universidad al Habla (UAH - University on the Line) spoken dialogue system is described. Firstly, in Section 2.2 there is a detailed description of the architecture of the system describing its main modules. The design principles followed are described, including efficient database access based on the interaction context, dynamic dialogue managing, tailored help messages and context-based oral responses. Special attention is paid to the creation of grammars with vocabulary which is previously unknown. Section 2.3 proposes a novel approach for the creation of speech recognition grammars from databases. This technique permits a reduction of the time employed during the speech recognition process to obtain a rapid system response to user prompts and also improves user satisfaction, as interaction is smoother.

The development of the UAH system was carried out not only to design and implement different dialogue management strategies in a real dialogue system, but also to employ a real spoken dialogue systems as a testbed in which to evaluate the different methodologies proposed in the Thesis. The system has been available to the public on the phone from June 2005, so that the users can interact with UAH to find out information related to the Dept. Languages and Computer Systems. A corpus of 422 user recordings corresponding to a year of UAH use was compiled and annotated, and was used for the approaches presented in the rest of the Thesis.

2.2 Modular architecture

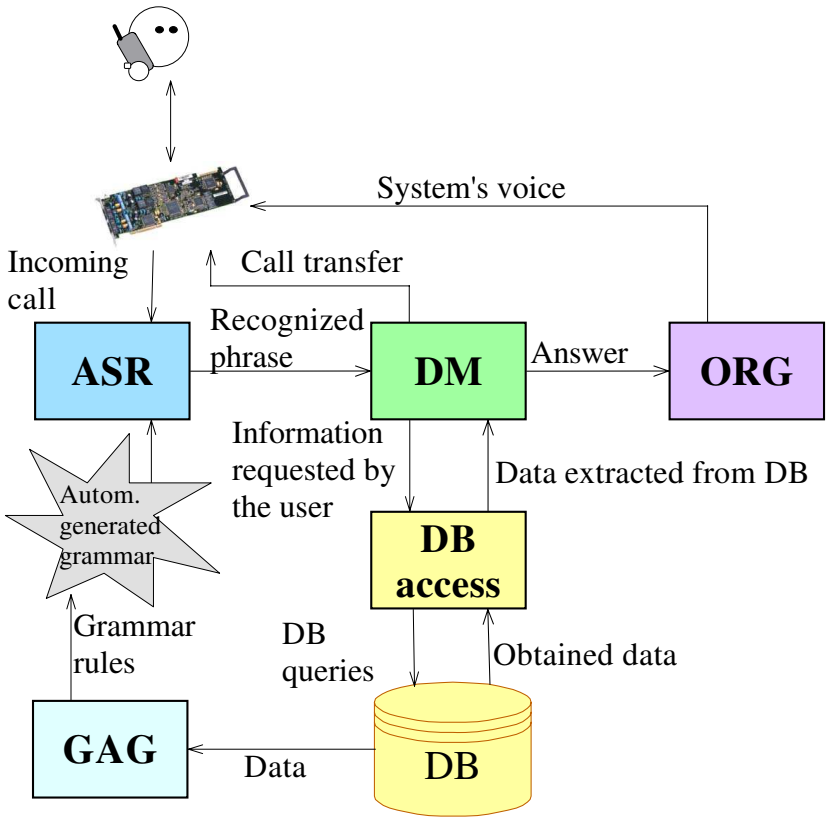


Figure 2.1. *Modular architecture of the UAH system*

UAH is a spoken dialogue system developed to provide spoken access to academic information of the Dept. Languages and Computer Systems, as well as some other additional information relating to the University of Granada. The first functional version of UAH was made available to the public in June 2005. Figure 2.1 shows the modular architecture of the UAH dialogue system. As can be observed, it is comprised of the five typical modules of current dialogue systems, concerned with automatic speech recognition (ASR), dialogue management (DM), database access (DB Access), data storage (DB) and oral response generation (ORG). In addition, there is a new module called GAG (Grammar Automatic Generation) which permits the automatic creation of the rules of grammars for speech recognition, which will be discussed in Section 2.3.

According to the data extracted from the user utterances, the dialogue manager (DM) decides the next system response. The dialogue manager is based on VoiceXML documents that are dynamically created using PHP as the interaction flows. UAH was designed to provide information about teachers, subjects, registration processes and post-graduate studies. In these scenarios different dialogue initiatives were used to obtain measures about how the interaction flexibility affects both the performance of the dialogue and user satisfaction. Information on teachers and subjects had to be accessed using a system-directed initiative, whereas dialogues providing information about registration and post-graduate studies were mixed-initiative.

The UAH system's dialogue manager dynamically adapts the system responses to the context and the dialogue state, for example to decide the confirmation strategy to be used. The system uses explicit confirmations for important actions or situations in which the ASR process is more difficult. For example, before transferring a call to a teacher the system explicitly confirms his/her name (e.g. "Did you say Mr. Smith?"). In the rest of the situations, the system uses explicit confirmations when confidence values are not high, and implicit confirmations otherwise, in order to provide the user with a more natural interaction. The system's oral response generation is carried out firstly in text format instantiating patterns. The phrases are tailored to the interaction context. For example, the system's help messages takes into account the topic the user and the system are talking about at a particular moment. Once the response is obtained in text format, it is transformed into voice using a commercial TTS system.

For purposes of illustration, we give below a sample dialogue translated from Spanish into English (S = system, U = user):

S1> Welcome to the UAH system. How can I help you?
U1> I need information about the registration procedure in the University.
S2> What do you want to know about the registration procedure?
U2> The deadline.
S3> The registration deadline is on the 31st of August. Do you need any other information about the registration procedure?
U4> No, thanks.
S5> Thank you for using the UAH system. Have a nice day.

2.2.1. Automatic speech recognition

The ASR module processes each user utterance provided by an Intel Dialogic D/41JCT-LS telephony card. This card handles the user call and gives the voice signal as a result, which is transformed into a word sequence in text format by the ASR module. The ASR process is carried out using several grammars that represent the valid phrases the users may utter. These grammars are created in different ways depending on the vocabulary and the moment of creation. In total there are four creation methods considering the combination of: i) known or unknown vocabulary at design time, i.e. never changing vocabulary, or vocabulary that can be constantly updated, as for example if it is stored in a database, and ii) static or dynamic grammar creation.

Static grammars with previously known content were created during the design of the dialogue system and do not vary their content with its operation. Their content is previously known and was carefully designed. These grammars are used in the UAH dialogue system for static menus that explain the information that the user can ask about. The UAH system also uses built-in grammars for the recognition of numbers, and Boolean answers like yes/no/true/false. Static grammars with unknown vocabulary will be described in Section 2.3.

Dynamic grammars are used in two ways. On the one hand, they can have a previously known content, which is dynamically generated in execution time only when they are really going to be used. On the other hand, UAH also employs grammars which are created dynamically from the vocabulary extracted from databases. This vocabulary is continuously changing and thus is not known at design time. This approach is used only when grammars have a small vocabulary, because creating grammars from large vocabularies at execution time would imply a delay in the interaction. This type of grammar is used, for example, in disambiguation grammars.

2.2.2. Dialogue management

The dialogue manager (DM) decides which system response to follow according to the data extracted from the user utterances. The complexity of this module depends on various factors. One is complexity of the modelled interaction, while another is the kind of task to be carried out by means of the

dialogue. The DM complexity depends as well on the flexibility desired for the dialogue and the type of initiative implemented (user directed, system directed or mixed). In the case of the UAH system, the application domain is information extraction in the University environment. The dialogue initiative is system-directed in the so-called basic election dialogue states, while it is mixed in the detail election states. In the basic election dialogue states the user decides, without going into detail, the kind of information s/he wants to know (e.g. information about teachers, subjects, timetables, etc.). In this case, the dialogue initiative is system-directed in order to ensure the user obtains the required information easily from the diverse kinds of information the system provides. On the other hand, in the detail election dialogue states, the user identifies the specific information s/he wants to obtain inside the previously selected area (e.g. a teacher's telephone number). The use of mixed initiative in these states provides the user with greater expression flexibility once the area has been determined (Narayanan et al., 2000). The system can take the initiative even in a detail election state if it detects a misunderstanding. In this case it provides the user with additional means for introducing the information. For example, if the user makes more than two mistakes in giving a teacher's full name, the system takes the initiative and gives prompts for the data one by one, i.e. first name and then surnames[1].

Although there is no consensus about the tasks a dialogue manager should perform, one of the most widely accepted proposals is presented by Traum and Larsson (2003). They propose that the dialogue manager must update the dialogue context to obtain the correct semantic interpretations from the user utterances. Furthermore, the dialogue manager must carry out tasks within a specific domain (in this case, the University context) and decide which information to provide to the user and when and how to express it.

The UAH dialogue manager dynamically adapts the system responses to the context and the dialogue state, changing some phrases to improve the naturalness of the interaction. For example, the system's help messages take into account the topic the user and the system are talking about at a particular moment.

[1]In Spain people have two surnames: father's and mother's

The context is used as well to decide the confirmation strategy to use. The system uses explicit confirmations for important actions or situations in which the ASR process is more difficult. For example, before transferring a call to a teacher the system confirms explicitly his/her name (e.g. "Did you say teacher Ms. Zoraida?"). In the rest of the situations the system uses implicit confirmations, in order to provide the user with a more natural interaction (Bernsen et al., 1994).

2.2.3. Database access

In the literature it is possible to find multiple references to the importance of separating the access and query of databases from the rest of the tasks carried out in a dialogue system. For example, in the GEMINI project (Hamerich et al., 2004), an assistant was built to connect to the database, so that users could create dialogue systems semi-automatically, regardless of the features of the database employed. A dichotomy between dialogue management and information access is achieved by means of creating a database access module. The dialogue manager supplies to this module the information that the user wants to know. The access module then constructs the query and extracts the data from the database.

Once the information is extracted, the access module executes a PHP program that validates the data. Besides, it checks that there are no repeated data in the result before sending it to the dialogue manager. Finally, the dialogue manager decides how to communicate the data to the user. However, there are situations in which the access module receives not only the kind of information to be extracted (e.g. telephone number of teachers called "Jones"), but also additional restrictions, such as gender. For example, if the user asks for Mrs. Jones's telephone number, then only data of female teachers with this surname are extracted.

The UAH system makes two types of database queries: explicit and implicit. In the first case, the query is carried out by the user initiative. For example, if the user asks for the telephone number of the teacher "John Jones", then two queries are executed. The first one checks the number of records in the teacher's database that correspond to the teacher's name provided. If the number is zero, the system informs the user no teachers were found with this name. If the number is greater than one, the system requests

the user to provide additional information to select the correct record. Once the name is determined, another query takes place to extract the telephone number. The first query is implicit because it is executed by the system's initiative, without an explicit user request. On the other hand, the second is explicit because it corresponds to the user request (in this example the user asked for a telephone number).

Furthermore, the complexity of the queries varies with the flexibility of the current dialogue state. In fact, queries vary from a simple "select FIELD from TABLE" to a very complex data selection based on partial or complete string matching. For example, when a user utters a teacher's name, the information provided can be a combination of name and surnames in up to seven different possibilities (e.g. name, name and first surname, first and second surname, etc.), including incomplete names (e.g. a user could say "Mary" instead of "Mary Kate").

The system extracts information querying a database which contains public information about the University of Granada (e.g. fax number and email of teachers, but not their personal data). The database has been designed to store data typically included in the departments' web pages. The UAH system could work with a database that stored many different kinds of information, creating views, if necessary, to keep economic and personal data private. The database used in the system's implementation is relational and has 22 tables that have been obtained from an entity-relation diagram, followed by a fusion and normalization process.

2.2.4. Oral response generation

The system's oral response generation is carried out firstly in text format using patterns which have been classified into several categories: teacher, PhD, subjects, registration, additional information, confirmation, greetings and help. Once the response is obtained in text format, it is transformed into voice using the Verbio TTS commercial system.

Inside each response category there is a specific pattern for each data the system could give to the user. These patterns are composed of several information segments that can be selected or ignored dynamically depending on the information to be provided. For example, the pattern used to inform about the location of a teacher's office is composed of three main information

segments: number of the office, name of Faculty where the office is, and floor inside the Faculty. If one of these data is not available at the current dialogue state, it is not included in the resultant text. For example, if the floor is not available the system response could be "office number 3 in Computer Science Faculty".

The patterns' morphological structure can also change according to the information provided, as it must take into account gender (e.g. "Ms. Mary Jones", "Mr. John Williams") and number (e.g. "The subject Parallel Programming has two credits", "The subject Software Engineering has one credit"). The system also uses a dynamic adaptation of linking words (e.g. "There are two teachers called Jones: Mary Jones and Michael Jones", "There are three teachers called Jones: Mary Jones, Michael Jones and William Jones").

Finally, the last stage in the creation of the system response in text form is the adaptation of several special words and symbols, as well as the inclusion of tags the TTS module uses to generate special pronunciations. The special words' adaptation translates the information stored in the database into a more suitable form to be read aloud by the system, whereas the tag inclusion permits a better understanding of the information synthesized by the system. For example, it is better to spell a web page's URL than reading it as if it were a long word. Also, it is preferable to read aloud a date in the format month-day-year than reading it.

2.3 Automatic grammar generation

Most of the commercial development environments for setting up spoken dialogue systems provide tools for rule-based grammar creation and testing. In these tools the grammar creation is done statically; that is, grammars are created in design time, before the system is put into service. Although this is the most straightforward way of creating grammars, static creation can cause inconsistencies between the database contents and the vocabulary included in the grammars when the vocabulary is not previously known or does not remain unchanged.

To overcome this drawback, there are several techniques for dynamic grammar creation. The first constructs the grammars dynamically in execution time (Truillet et al., 2004). This technique offers flexibility as the

grammars are always updated with the last changes in the database. Nevertheless, this method may imply a very high computational load, which introduces a delay time that, in systems with very large databases (e.g. academic information of a University), can be excessive; thus causing the system to be considered "slow" by users.

A second technique creates the grammars at the beginning of each interaction, before the ASR process takes place. These grammars are always updated regardless of changes in the database (Schalkwyck and Story, 2003). This method does not imply an increase in the ASR time. There is therefore less increase in waiting time from the user point of view, but there is still a delay in the system start-up that increases with the database size.

In both cases, the suitability of the technique depends on the size of the vocabulary. As has been shown previously (Nielsen, 1994), there are three basic execution times to be considered. Firstly, the limit for the user to consider the interaction as real-time is 0.1 seconds. Secondly, from 0.1 to 1.0 seconds, which is the interval in which the user notices the delay but his flow of thought is uninterrupted. And thirdly, from 1.0 to 10.0 seconds, when the user still pays attention to the system. However for delays longer than 10 seconds it is necessary to give some feedback to the user with a system prompt and/or music (Cerrato, 2002).

The execution time needed to construct a grammar rule from a specific column in a database was measured regarding the number of words gathered. As shown in Figure 2.2, the 0.1 limit is reached with 10,000 words, the 1 second limit is reached with 100,000, and the maximum limit (10 seconds) is reached with vocabularies around 300,000 words. For example, for a vocabulary of 1 million words there is a 30 seconds delay. Thus, for large vocabularies (more than 300,000 words) the overload introduced is noticed by the user and a feedback technique is needed. However, feedback techniques usually cause a bad impression on users, who usually find them very irritating (Mäkelä et al., 2001). Furthermore, a minimal system response delay (defined as the time elapsed between the user finishing talking and the system response (Möller, 2005)) is vital to minimize interaction costs. This, coupled with the maximization of task success, is essential to enhance user satisfaction and is usually employed as a key parameter for spoken dialogue systems evaluation, as proposed in the PARADISE framework (Walker et al., 2000a).

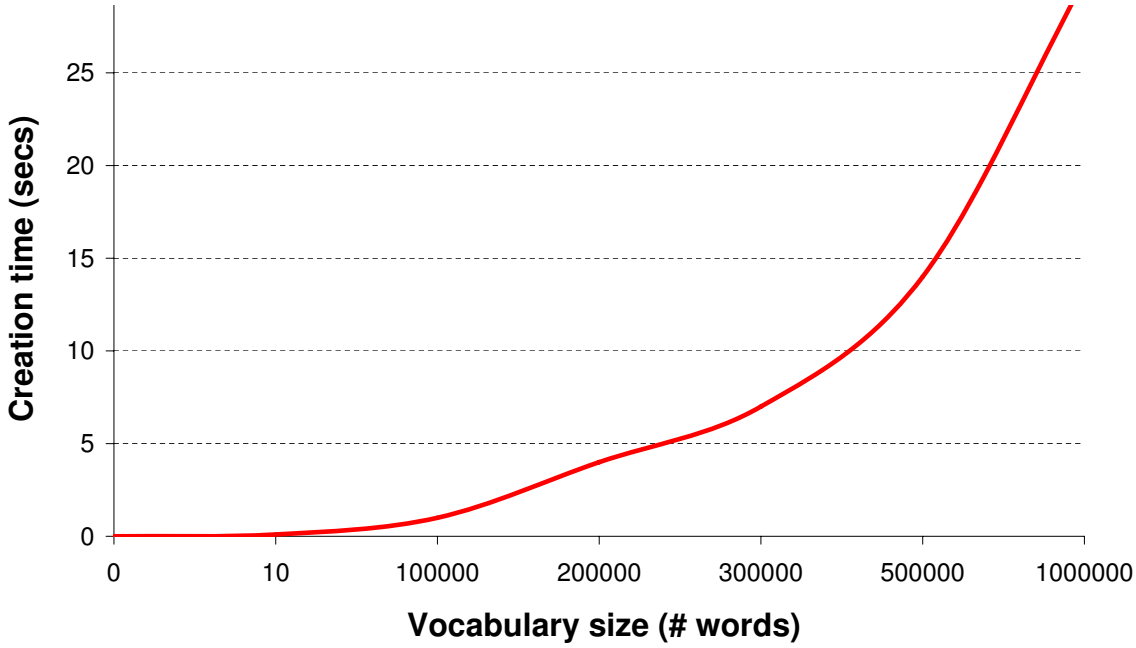


Figure 2.2. *Latency of the dynamic creation of grammar rules*

To avoid the problems discussed above, the TGC technique (Triggered Grammar rules Creation) can be employed. TGC uses automatically-generated static grammar rules which are updated using a triggering mechanism as the database changes. This way there is no creation delay in execution time and the user feels he is interacting with a fast system.

The main advantage of the TGC technique in comparison with others we can find in the literature (e.g. McTear (2004)) is that it is general and thus suitable for generating grammar rules to be used by any dialogue system that extracts vocabulary from a database.

The creation of grammar rules is carried out following three steps: i) information is extracted from the database, ii) the grammar rules are created from the extracted data formatting it into JSGF (Java Speech Grammar Format) or ABNF (Augmented Backus-Naur Form); in this way, grammar rules are generated before the ASR is carried out; iii) grammar rules are updated when the database changes. To ensure grammars are permanently updated when the database changes, the proposed technique employs a mechanism based on database triggers. These are fired when changes occur in the database fields from which the vocabulary of the grammar rules is extracted.

The W3C Voice Browser working group specifies in (Brown, 1999) that rules must be re-definable in execution time. To do so, they point out several mechanisms, as for example the division of the rules space into static and dynamic arenas. Using the proposed technique, the sentence structure is created statically containing references to dynamic rules, which can be stored in the same file or in an external one. The TGC technique updates these rules with new data extracted from the database without affecting the syntactic structure of the phrases.

The TGC technique was implemented in the so-called Grammar Automatic Generation (GAG) tool using PHP, HTML, JavaScript and PostGreSQL. This tool is used as an additional module the UAH system. This tool is used as an additional module in the UAH system and has an easy-to-use web interface to let the system designer choose the database fields to be used to extract the vocabulary. The process of creating rules using the tool is divided into three steps. Firstly, the GAG tool prompts the user to provide the database name, host, user name and password to access the database where the information is stored. It also prompts for the name of the grammar rule and the type of grammar to be created (JSGF or ABNF). Secondly, the selection of fields from the database tables is visualized in a drop-down menu. Fields appear in the menu in the same order as in the corresponding table, but the system designer can select them in any other to create a grammar rule. When a field is selected, a number appears automatically next to the field indicating the selection order. For example, in an academic application the system designer could select a teacher's name and surname, obtaining the grammar rule <teacher> = ("Claire Smith" | ... | "Jack McNeal"). Alternatively, he could select for example the teacher's surname and name, in which case the obtained grammar rule would be: <teacher> = "Smith Claire" | ... | "McNeal Jack"). Thirdly, after the field selection the designer must enter the file name to store the grammar rule. If he marks the "Add to file" option in the interface, the rule is created as a part of an existing grammar. Otherwise, a new grammar with just one rule is created in the specified file.

The GAG tool implements a triggering mechanism to keep the grammar rules updated with the last changes made in the database. To do so, at the end of each grammar creation the system designer chooses whether to update automatically the vocabulary in the rule with database changes. If

he indicates the rules must be automatically updated, triggers are dynamically created to be activated when the values of the corresponding columns are updated, deleted or inserted. For example, if he creates a grammar rule for teachers' names and surnames, the trigger includes "obligatory" as a new word in the rule if this word is in the field "Type" of the "Subject" table. It also deletes the word "obligatory" from the rule if it is erased from the table, and changes "obligatory" to "optional" if the field is updated with the latter word.

As triggers provide the old and new values of the table field, only the pertinent changes are made in the grammar rules instead of generating them again right from the start when any change takes place, which introduces a delay equivalent to creating them dynamically. The whole process is illustrated in Figure 2.3, where a distributed architecture is depicted in which the ASR takes place in a computer with a telephony card, while the dialogue manager and the GAG tool operate in an application server that uses the data provided by a server in which the database is hosted.

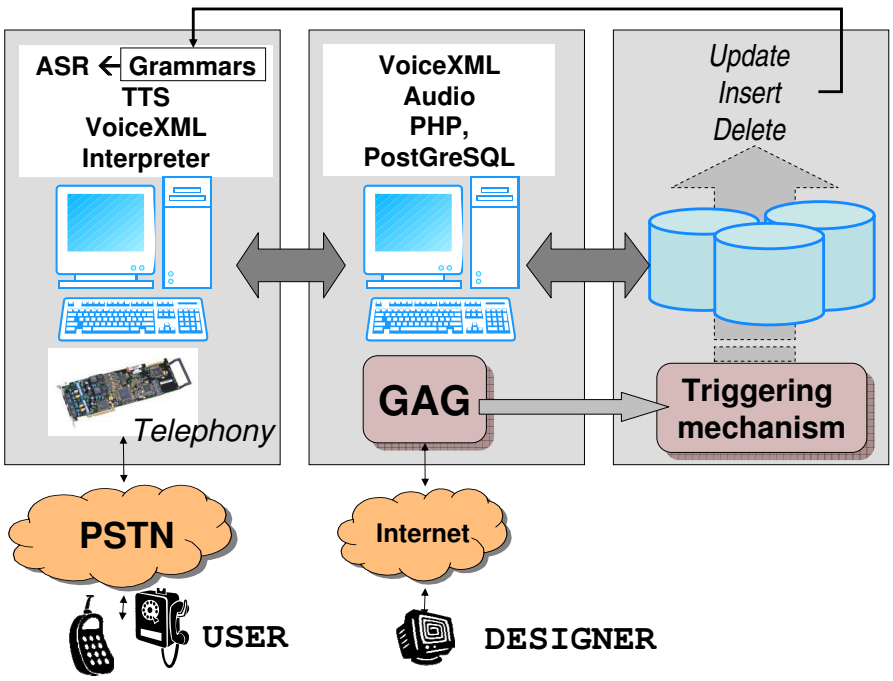


Figure 2.3. *Automatic grammar rules update with the GAG tool*

Two different approaches to addressing the creation of ASR grammar rules from databases were tested: dynamic creation and creation using the GAG module. The experiments were carried out employing different vocabulary sizes (as shown in Figure 2.2). Figure 2.4 shows that user satisfaction with dynamic grammar creation decreases with larger vocabularies, which indicates that user preference for the GAG tool outperforms preference for dynamic creation in most cases. There is a correspondence between user satisfaction and the temporary constraints previously commented.

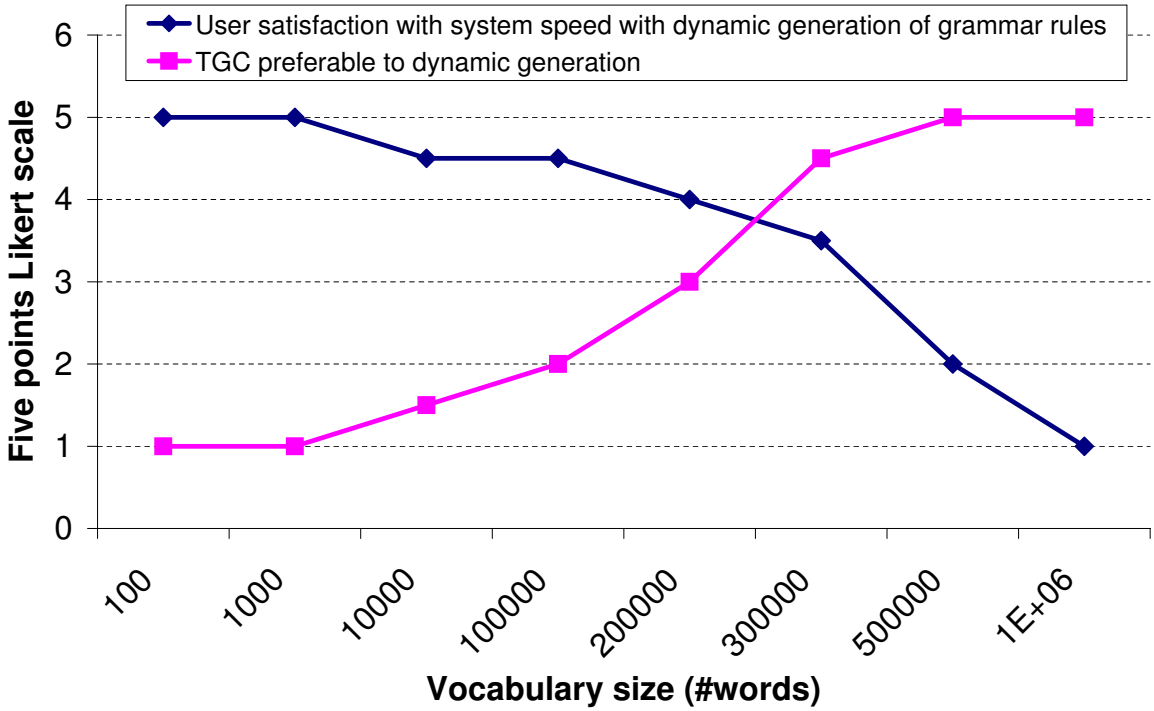


Figure 2.4. *TGC technique vs. dynamic creation of grammar rules measured in terms of user satisfaction*

Thus, in vocabularies with less than 10,000 words (generation time less than 0.1 seconds) users are not aware of the system delay with the dynamic technique, which makes both techniques be equally rated by users. It is also observable that in this figure there is a point at which both lines cross. This corresponds to the value around 300,000 words for which a vocabulary can be considered large enough to make the proposed technique clearly preferable to dynamic creation, given the delay introduced by the latter and its consequent negative effect on users' satisfaction.

Furthermore, regarding the use of the GAG tool, it is shown that the users' perception varies depending on the type of interaction initiative used by the dialogue system. Although 71% of the users find the interaction speed adequate or fast, 20% of them consider it slow (only 3% thought it very slow). This occurs because of the system-directed initiative, which presents several options to the user in order to obtain the kind of information he wishes. As shown in Figure 2.5, overall user satisfaction with the interaction (rated on a five-point Likert scale) is in 78% of cases greater than or equal to 3, which shows the good results achieved by the system employing the proposed technique.

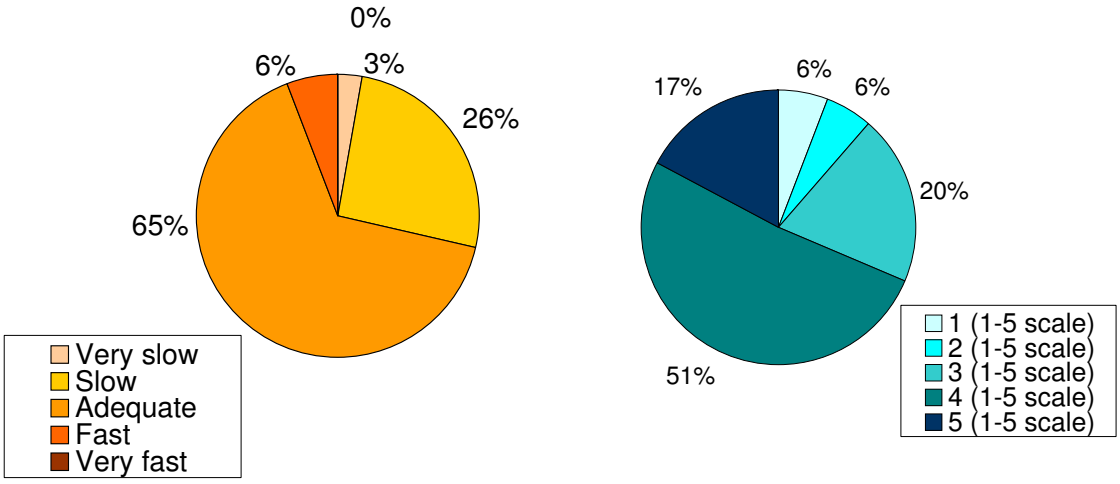


Figure 2.5. *Perceived interaction speed (left) and user satisfaction (right) using UAH*

It was found that, even when the objective evaluation measures indicate that the system delay is negligible, users sometimes are not of the same opinion. Thus, users may consider the interaction slow if the initiative is system-directed, even when the system response time is small. The reason for this is that they feel that the interaction could be faster if they had the opportunity to directly utter their queries.

2.4 The UAH speech corpus

The corpus used for the experiments described in this Thesis is comprised of 85 dialogues by 60 different users interacting with UAH. It contains 422 user turns, with an average of 5 user turns per dialogue. The recorded material has a duration of 150 minutes. It was semi-automatically annotated with de-facto standard evaluation criteria and the users' opinions were obtained with questionnaires as described in Chapter 5. Each user utterance was subsequently tagged with an emotional state by nine non-expert annotators as described in Chapter 3. The size of the corpus is similar to that of other real emotional speech corpora such as those used by Forbes-Riley and Litman (2004a) (10 dialogues, 453 turns) or Morrison et al. (2007) (391 user turns).

2.5 Conclusions

UAH is a spoken dialogue system developed in VoiceXML standard language. It uses dynamic dialogue managing and introduces several novel techniques that make the interaction smoother. It was developed to cover different interaction initiatives and confirmation strategies in order to be able to compare their suitability. The introduction of the GAG module is noteworthy. This provides a new way of creating rules for speech recognition grammars without introducing delays and keeps the vocabulary updated. The UAH system has been functional since 2005 when it was made available to the public. Since then, the interactions are being recorded and the main interaction parameters are automatically processed to compute evaluation criteria. The dialogues corresponding to a year of usage of the UAH system were also annotated to introduce new parameters and the opinions of the users about the system have been surveyed. With this information a speech corpus was created which formed the basis for the research described in the next chapters.

Innegable, señor.
Es indisimulable.
¿Está usted aburrido?
Me parece que está usted aburrido.
Dígame, ¿adónde va tan aburrido?

Rafael Alberti, El aburrimiento

3 Recognition of non-acted emotions

3.1 Introduction

One of the main research objectives of dialogue systems is to achieve human-like communication between people and machines. This eliminates the need for keyboard and mouse in favour of more intuitive ways of interaction, such as natural language, thus leading to a new paradigm in which technologies can be accessed by non-expert users or handicapped people.

However, multimodal human-computer interaction is still not comparable to human dialogue. One of the reasons for this is that human interaction involves exchanging not only explicit content, but also implicit information about the affective state of the interlocutor. Systems that make use of such information are described as incorporating "affective computing" or "emotion intelligence", which covers the areas of emotion recognition, interpretation, management and generation.

Due to its benefits and huge variety of applications, affective computing has become an outstanding research topic in the field of HCI, and numerous important international and interdisciplinary related projects have appeared. Some of the latest are MEGA (Camurri et al., 2004), NECA (Gebhard et al., 2004), VICTEC (Hall et al., 2005), NICE (Corradini et al., 2005), HUMAINE (Cowie and Schröder, 2005) and COMPANIONS (Wilks, 2006), to mention just a few.

Accurate annotation is a first step towards optimized detection and management of emotions, which is a very important task in order to avoid

significant problems in communication, such as misunderstandings and user dissatisfaction, that lead to very low task completion rates. Despite its benefits, the annotation of emotions in spoken dialogue systems encounters restrictions as a result of certain important problems. Firstly, as shown in Section 3.3.2, the percentage of neutral vs. emotive speech is usually very unbalanced (Forbes-Riley and Litman, 2004a; Morrison et al., 2007). Secondly, all information must be gathered through the oral modality and in some systems where the dialogue is less flexible, the length of the user utterances can be insufficient to enable other knowledge sources like linguistic information to be employed.

To solve these problems, we propose to use contextual information for the annotation of user emotions in spoken dialogue systems. The main interest is to recognize negative emotions as some studies, like for example (Riccardi and Hakkani-Tür, 2005), have shown that once the user is in a negative emotional state, it is difficult to guide him out. Furthermore, these bad experiences can also discourage users from employing the system again. Concretely, three negative emotions are taken into account. The first is *doubtful*, which is useful to identify when the user is uncertain about what to do next. A user is in this emotional state when he has doubts about what to say in that turn. The second and third emotions are *angry* and *bored*, two negative emotional states that must be recognized before the user gets too frustrated because of system malfunctions. In the activation-evaluation space (Russell, 1980; Scherer, 2005), *angry* corresponds to an active negative emotion, whereas *bored* and *doubtful* to passive negative emotions.

Different approaches are presented in this chapter in order to include contextual information in both human annotation (as discussed in Section 3.3) and machine learned classification (Section 3.4). In human annotation, non-expert annotators were provided with contextual information by giving them the utterances to be annotated along with the dialogues where these were produced. In this way, annotators had information about the user speaking style and the moment of the dialogue at which each sentence was uttered. For machine-learned classification, a novel method in two steps is introduced, which enhances negative emotion classification with automatically generated context information. The first step calculates users' neutral speaking style, which is employed to classify emotions into *angry* and *doubtfulORbored*; whereas the second step introduces dialogue context and

allows the distinction between *bored* and *doubtful* categories. One of the main advantages of the proposed method is that it does not require including additional manually annotated data. Hence, it permits a straightforward automatic integration of context information within emotion recognizers for spoken dialogue systems.

To evaluate the benefits of the proposals, different experiments have been performed over a corpus of real emotions extracted from the interaction of 60 different users with the UAH spoken dialogue system (Chapter 2). The objective is to demonstrate that the proposed contextual information influences human as well as machine recognition, and that better results can be obtained when context is included using the proposed methods, if compared to recognition based on traditional acoustic features or the baseline classification methods.

The chapter is structured as follows: in Section 3.2 there is an overview of the related work done in the area and the points in which the Thesis makes its main contributions. Section 3.3 presents the human annotation procedure and discusses the corpus facts and the results in terms of emotions annotated and agreement between annotators. In Section 3.4 there is a description of the automatic classification of emotions, and a discussion of the experimental results obtained for it. Section 3.5 describes the previous approaches studied to take into account both context sources until the optimal approach was encountered. Finally, in Section 3.6 there is a summary of the benefits of the proposed methods and present the conclusions extracted from them.

3.2 Related work

Unfortunately there is no consensus about what an emotion is. Psychological and biological studies in this matter have been carried out for centuries and it is not strange to find references to Darwin or Descartes in some recent papers about the topic. In an effort to clarify concepts and underlying the direction of main research lines in the area, Cowie (2000) distinguishes two senses in which the word emotion can be interpreted: the first one is as discrete states (p.e. fear, happiness, anger) which are usually referred to as "full-blown" emotions in the literature; and the second one is as an attribute of certain states, which the author names as "emotional states". In the HUMAINE project Humaine emotion-research.net (2007) they also make

this distinction using other nomenclature: episodic and pervasive emotions. Main research efforts are carried out towards the study of pervasive emotions or "emotional states"; whereas the main objective in the study of "full-blown" emotions is to find a restricted set of categories or emotions. There exists the extended theory that "full-blown" emotions can take only a few forms easily distinguishable from each other. However, making such a simplification is not always suitable as in some applications it can be interesting to study blended, simulated and/or conflictive emotions.

Regardless of the definition of emotion, several ways to represent them can be found in the literature. They can be represented using a discrete set or as points in a continuous space. In the continuous case, emotions are represented by coordinates in a space with a small number of dimensions. The typical approach is the bidimensional activation-evaluation space Cowie et al. (2001). In the horizontal axe, evaluation deals with the "valence" of emotions, that is, positive or negative evaluations of people, things or events. In the vertical one, activation measures the user disposition to take some action rather than none. "Full-blown" emotions form a circular pattern in the activation-evaluation space which made other authors propose a representation based in terms of angles and distance to the centre. Some tools like for example FEELTRACE (Cowie et al., 2000) have been implemented to give a visual representation of the dynamic progress of emotions inside this circle. Additionally, 3D models can be used to distinguish between emotions that are very near each other in the circle (e.g. fear and anger), the new dimensions used are usually perceived control or inclination to engage. Emotions can also be represented in a structural way, which treats them from a cognitive perspective that describes how users deal with the situation that caused the emotion. Furthermore, following the Ortony and G. L. Clore (1988) theory, emotions can be classified according to the emotion-eliciting situations which can be related to events, actions of agents or aspects of objects. OCC theory is usually employed for emotion synthesis (Zong et al., 2000; de Melo and Paiva, 2005).

Emotionally intelligent systems can put the emphasis on in the emotions' cause or effect. In the first case the focus is in the reasons of the apparition of some emotion, which can be external or internal to the user; the second describes the effects of these characteristics in the listener Cowie (2000). The research in these areas is generally recognition driven in the

cause-type case and synthesis driven in the effect-type case. The research described in this chapter is focused on the emotion recognition case.

Emotion recognition can be carried out with invasive and non invasive methods. Invasive methods are based in physiological measures like breathing rate or conductivity of skin Picard (1997). One of the most widespread methods consists in measuring the galvanic skin response (GSR) as there is a relationship between the arousal of emotions and changes in GSR Lee et al. (2005). Some other methods are EMG, which measures facial muscles Mahlke (2006), hear rate or more recently the usage of brain images Critchley et al. (2005). Non invasive methods are usually based in audio and video. On the one hand, audio emotion recognition can be carried out from the acoustic information or from linguistic information. Speech is deeply affected by emotions: acoustic, contour, tone, voice quality, articulation change with different emotions, a comprehensive study of those changes is presented in Cowie et al. (2001). Language information deals with linguistic changes depending on the emotional state of the user. For this purposes the technique of word emotional salience has gained remarkable attention. This measure represents the frequency of apparition of a word in a given emotional state or category and it is calculated from a corpus of user-system interactions Lee et al. (2005). On the other hand, video recognition usually pays attention to facial expression, body posture and movements of the hands; a summary of all these features can be found in Picard and Daily (2005). Other authors emphasize that emotions are influenced by cultural and social settings and defend an "interactional approach" Boehner et al. (2007) to be considered along with physiological, audio or video measures.

Emotion recognition has been used in Human Computer Interaction (HCI) systems for several purposes. In some application domains it is necessary to recognize the affective state of the user to adapt the systems to it or even change it. For example, in emergency services (Bickmore and Giorgino, 2004) or intelligent tutors (Ai et al., 2006), it is necessary to know the users' emotional state to calm them down, or to encourage them in learning activities. However, there are also some applications in which emotion management is not a central aspect, but contributes to the better functioning of the system as a whole. In these systems emotion management can be used to resolve stages of the dialogue that cause negative emotional states, as well as to avoid them and foster positive ones in future interactions. For exam-

ple, Burkhardt et al. (2005) use an anger detector to avoid user frustration during the interaction with their voice portal. Furthermore, emotions are of interest not just for their own sake, but also because they affect the explicit message conveyed during the interaction: they change peoples' voices, facial expressions, gestures, speed of speech, etc. This is usually called "emotional colouring" and can be of great importance for the interpretation of user input. For example, Wahlster (2006) use emotional colouring in the context of the SmartKom system to detect sarcasm and thus tackle false positive sentences.

As explained before, emotion recognition is a key aspect to obtain human-like interaction. That is why it has received a lot of attention for the dialogue systems research community. From applications in with the changes in the users' emotional state are uniquely indicators on when the system is not fulfilling users' expectations, to complicated systems in which emotions are a key-stone, like psychological aid systems; emotion recognition is gaining increasing attention form the research community. This is reflected in the number of international interdisciplinary projects that have treated the topic, some of the latest are:

- SAFIRA - Supporting Affective Interactions for Real-time Applications (The Safira Project - DFKI Page, 2002). 24 months from 2000-05-02 (Completed). Its purpose was the enrichment of applications with an affective dimension to support affective behaviour and control in real-time multi-agent systems interacting with users.

- MEGA - Multisensory Expressive Gesture Applications (MEGA, 2001). 36 months from 2000-11-01 (Completed). Its purpose was the modelling and real-time analysis, synthesis, and networked communication of expressive and emotional content in non-verbal interaction (e.g. music, dance) by multi-sensory interfaces, from a multimodal perspective.

- MAGICSTER - Embodied Believable Agents (MagiCster Project Pages, 2007). 39 months from 2000-12-01 (Completed). Its purpose was the design and evaluation of a believable conversational interface agent, which makes use of gaze, facial expression, gesture and body posture as well as speech in a synchronised fashion.

- NECA - A Net Environment for embodied emotional Conversational Agents (NECA Project, 2005). 30 months from 2001-10-01 (Completed). Its purpose was the creation of multi-user and multi-agent virtual spaces populated by affective conversational agents able to express themselves through synchronised emotional speech and non-verbal expression.

- ERMIS - Emotionally Rich Man-Machine Interaction Systems (EUROPA - CORDIS: Community Research and Development Information Service, 2006). 36 months from 2002-01-01 (Completed). Its purpose was the development of a prototype system for human computer interaction that can interpret users' attitude or emotional state, in terms of their speech and/or their facial gestures.

- PF-STAR - Preparing future multisensorial interaction research (PF-STAR home page, 2004). 24 months from 2002-10-01 (Completed). Its purpose was the contribution to the field of multisensorial and multilingual communication by providing technological baselines, comparative evaluations, and assessment of prospects of core technologies specially in the topic of technologies for speech-to-speech translation, the detection and expressions of emotional states, and core speech technologies for children.

- VICTEC - Virtual ICT with Empathic Characters (VICTEC in Lynne Hall web page, 2005). 35 months from 2002-03-01 (Completed). Its purpose was the development of a synthetic characters toolkit that supports the creation of believable synthetic characters in a virtual environment who establish credible and empathic relations with children.

- NICE - Natural Interactive Communication for Edutainment (NICE project - Main page, 2007). 36 months from 2002-03-01 (Completed). Its purpose was to foster universal natural interactive access, in particular for children and adolescents, by developing natural, fun and experientially rich communication between humans and embodied historical and literary characters.

- CHIL - Computers in the Human Interaction Loop (CHIL - Computers In the Human Interaction Loop, 2007). 36 months from 2003-12-18

(Completed). Its purpose was the creation of environments in which computers serve humans that focus on interacting with other humans instead of having to attend to and being preoccupied with the machines themselves.

- HUMAINE - Research on Emotions and Human-Machine Interaction (Humaine emotion-research.net, 2007). Since 2003-12-18 (In execution). Its purpose is to lay the foundations for European development of systems that can register, model and influence human emotional and emotion-related states coordinating efforts to come to a shared understanding of the issues involved.

- AMI - Augmented Multi-party Interaction (Augmented Multiparty Interaction Project, 2007). Since 2003-12-18 (In execution). Its purpose is the creation of new multimodal technologies to support human interaction in the context of smart meeting rooms and remote meeting assistants.

- INTREPID - A Virtual Reality Intelligent Multi-sensor Wearable System for Phobias' Treatment (Intrepid Project , A virtual reality intelligent multi-sensor wearable system for phobias' treatment). 24 months from 2004-01-01 (Completed). Its purpose was the development of a multi-sensor context-aware wearable system for the treatment of phobias.

- AUBADE - A wearable EMG Augmentation system for robust behavioural understanding AUBADE (2005). 34 months from 2004-01-01 (Completed). Its purpose was the development of a wearable platform, to ubiquitously monitor and recognise the emotional state of its users in real time, using signals obtained from their face.

- COSY - Cognitive Systems for Cognitive Assistants CoSy Home (2007). 48 months from 2006-09-01 (In execution). Its purpose is the construction of physically instantiated systems that can perceive, understand and interact with their environment, and evolve in order to achieve human-like performance in activities requiring context-(situation and task) specific knowledge.

- CALLAS - Conveying Effectiveness in Leading-Edge Living Adaptive Systems Callas - Conveying Affectiveness in Leading-Edge Living Adaptive Systems (2007). 42 months from 2006-11-01 (In execution). Its purpose is the definition and development of a multimodal architecture including emotional aspects, to support experiments and target new media applications essentially in an ambient intelligence paradigm.

- COMPANIONS - Intelligent, Persistent, Personalised Multimodal Interfaces to the Internet Companions (2007). 48 months from 2006-11-01 (In execution). Its purpose is the creation of companions: personalized, conversational interface to the Internet that knows its owner, on a range of platforms, indoor and nomadic, based on integrated high-quality research in multi-modal human-computer interfaces, intelligent agents, and human language technology.

In the area of emotion recognition the great majority of studies[1] focus on studying the appropriateness of different machine learning classifiers (Shafran and Mohri, 2005), such as K-nearest neighbours (Lee and Narayanan, 2005), Hidden Markov Models (Ververidis and Kotropoulos, 2006; Pitterman and Pitterman, 2006), Support Vector Machines (Morrison et al., 2007), Neural Networks (Morrison et al., 2007) or Boosting Algorithms (Liscombe et al., 2005; Forbes-Riley and Litman, 2004a). In addition, important research has been directed towards finding the best features to be used for classification. These features can be categorized at different levels. The lowest level deals with physiological features, which are usually measured with intrusive methods. Some examples are galvanic skin response (Lee et al., 2005), facial muscle movements (Mahlke, 2006) or brain images (Critchley et al., 2005). Acoustic and linguistic levels are more widespread and features like articulation changes (Cowie et al., 2001), statistical measures of acoustic features (Ververidis and Kotropoulos, 2006) or word emotional salience (Lee and Narayanan, 2005) are frequently found in the literature. In addition, visual features like facial expression, body posture and movements of hands have recently been adopted, especially in multimodal systems (Picard and Daily, 2005; Zeng et al., 2006). More recently, some authors like Boehner et al. (2007) have proposed cultural information as an additional source of information for detecting emotional states.

[1]For further information, see Ververidis and Kotropoulos (2006).

However, less attention is being paid to the training process of the algorithms in which automatic emotion classification is based, and for which emotional annotated corpora are needed. A good annotation scheme is essential as it affects the rest of the stages in the learning process. Besides, manual annotation of corpora is very difficult, time-consuming and expensive, and thus must be carefully designed. Authors that study emotional corpora are mainly interested in how it is gathered, especially comparing acted vs. real emotions acquisition (Morrison et al., 2007), but less work has been done in how the annotation of such a corpus must be achieved. Among others, Devillers et al. (2005) have proposed guidelines to design and develop successful annotation schemes in terms of labels, segmentation rules and validation processes. Gut and Bayerl (2004) have also worked on reliability measures of human annotations, whereas Craggs and Wood (2003) have proposed several layers of emotion annotation.

This Thesis goes a step further and studies how to add contextual information to the corpus annotation process, and suggest the inclusion of two new context sources: users' neutral speaking style and dialogue history. The former provides information about how users talk when they are not conveying any emotion, which can lead to a better recognition of users' non-neutral emotional states (Section 3.4.2). The latter involves using information about the current dialogue state in terms of dialogue length and number of confirmations and repetitions (as will be discussed in Section 3.4.3), which gives a reliable indication of the users' emotional state at each moment. For example, the user is likely to be angry if he has to repeat the same piece of information in numerous consecutive turns in the dialogue.

In the literature there are three main approaches for collecting emotional speech corpora: recording spontaneous emotional speech, recording induced emotions, and using actors to simulate the emotions. As shown in Figure 3.1, in these approaches there is a compromise between naturalness of the emotions and control over the collected data: the more control over the generated data, the less spontaneity and naturalness of the expressed emotion, and vice versa. Therefore, spontaneous emotional speech, which reflects completely natural emotional speech production in the application domain of the emotion recognizer, is the most realistic approach. However, a lot of effort is necessary for the annotation of the corpus, as it requires an interpretation of which emotion is being expressed in each recording. Sometimes,

the corpus is recorded from human-to-human interaction in the application context (Forbes-Riley and Litman, 2004a). In these cases, the result is also natural but it is not directly applicable to the case in which humans interact with a machine. In the other extreme, acted emotional speech is easier to manipulate and avoids the need for annotation, as emotions conveyed in each recording are known beforehand. The results obtained with acted speech are highly dependent on the skills of the actors, therefore the best results are obtained with actors with good drama preparation. When non-expert actors are used, another phase is necessary to discard the recordings that fail to reproduce the required emotion appropriately. In a middle point are the induced emotions, which can be more natural, like the ones elicited when playing computer games (Johnstone, 1996), or easier to manipulate like the ones induced by making people read texts that relate to specific emotions (Stibbard, 2000).

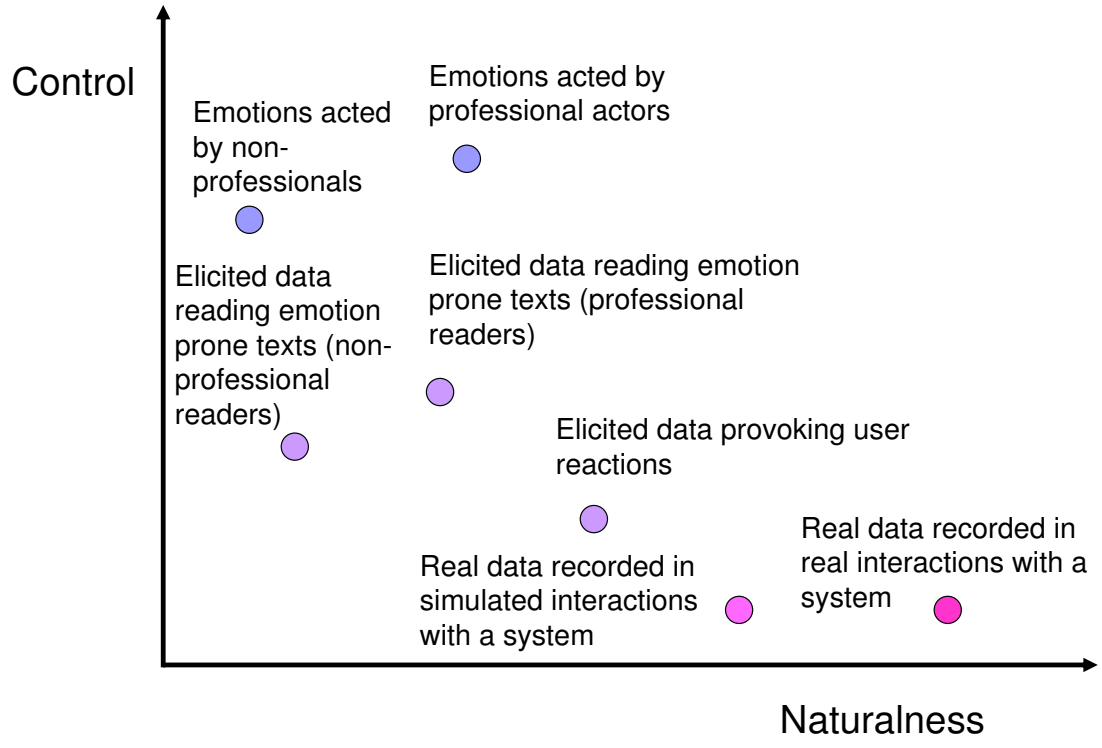


Figure 3.1. *Naturalness vs. control in the main emotional corpora generation approaches*

As some authors have indicated, e.g. Douglas-Cowie et al. (2003), the relationship between acted data and spontaneous emotional speech is not exactly known. But, as stated by Johnstone (1996), even professionally acted speech loses realism as there are some effects that cannot be controlled consciously. Thus, different studies have shown that it is not appropriate to use acted data to recognize naturally occurring emotions (Vogt and André, 2005; Wilting et al., 2006).

As the objective is to build an emotion recognizer for the UAH dialogue system (Chapter 2), and it would have to work with natural emotions occurring in real time, an utterance corpus collected from real users interacting with the system was used in the experiments. This is an important contribution to the state of the art as real non-elicited emotions are difficult to find in Spanish corpora. For example, out of the 70 corpora studied by Douglas-Cowie et al. (2003) and Ververidis and Kotropoulos (2006), only three are in Spanish: González (1999), Montero et al. (1999) and Iriondo et al. (2000). As shown in Table 3.1, two of these are used for emotion synthesis instead of recognition. The table also sets out Spanish corpora employed for emotion feature studies (Adell et al., 2005) and general purpose studies (Hozjan et al., 2002). None of these corpora were collected from real user interactions and the maximum number of actors used was 8, whereas the UAH corpus was collected from 60 users and real emotional speech.

Reference	Actors/users	Purpose	Kind
González (1999)	-	Recognition	Elicited
Montero et al. (1999)	1 actor	Synthesis	Simulated
Iriondo et al. (2000)	8 actors	Synthesis	Simulated
Hozjan et al. (2002)	2 actors	Study, synthesis and recognition	Simulated
Adell et al. (2005)	1 actress, 1 professional reader, 1 member of Spanish Parliament	Emotion features study	2 simulated, 1 natural
UAH corpus	60 UAH system users	Recognition	Natural

Table 3.1. *Summary of Spanish emotional speech corpora*

3.3 Human annotation of the UAH corpus

The annotation of emotions is a highly subjective task, given that for the same utterance, different annotators may perceive different emotions. The most reliable way to obtain rigorous annotations is to recruit specialized annotators, for example psychologists who are trained to recognize human emotions. Unfortunately, in most cases expert annotators are difficult to find and thus the annotation must be done by non-experts. In this case, all annotators were non-expert as they had not received any specific training on emotion recognition.

To get the best possible annotation employing non-expert annotators, the labelling process must be rigorously designed. Vidrascu and Devillers (2005) suggest several phases to decide the list of labels and annotation scheme, segmentation rules, number of annotators, validation procedures, and consistency study.

The first step is to decide the labels to be used for annotation. Our main interest is to study negative emotional states of the users, mainly to detect frustration because of system malfunctions. Thus, classification is made between the three major negative emotions encountered in the UAH corpus, namely *angry*, *bored* and *doubtful*. For the human annotation of the corpus a fourth category has been used: *neutral*, which represents a non-negative emotional state (i.e. positive emotions such as happiness are also treated as *neutral*). The neutral category was used only for the human annotation of the corpus. The rest of the experiments will focus exclusively on the distinction between the negative emotions considered.

A decision was made to use an odd, high number of annotators - nine, which is more than is typically reported in previous studies, e.g. Forbes-Riley and Litman (2004a) and Lee and Narayanan (2005). Regarding the "segment length", in this study this is the whole utterance because it was not useful to employ smaller segmentation units (i.e. words). The reason is that our goal was to analyse the emotion as a whole response to a system prompt, without considering the possible emotional changes within an utterance.

In the proposed annotation procedure the corpus was annotated twice by every annotator, firstly in an ordered style and secondly in an unordered style. In the first mode the annotators had information about the dialogue context and the users' speaking style. In the second case the annotators

did not have this information, so their annotations were based only on the acoustic information of the current utterance.

The final emotion assigned to each utterance in the ordered and unordered schemes was the one annotated by a majority of annotators in each of them. Gold standard emotions for the whole corpus were then computed from the results of each of the schemes. In situations where there was no majority of an emotion above the others (e.g. 4 *neutral*, 4 *bored* and 1 *doubtful*), priority was given to the non-neutral ones (in the last example *bored*). If this conflict was between two non-neutral emotions (e.g. 4 *doubtful*, 4 *bored* and 1 *neutral*), the results were compared between both annotation schemes to choose the emotion annotated by majority among the 18 annotations (the 9 of the ordered and the 9 of the unordered scheme).

3.3.1. Calculation of the agreement between annotators

Several Kappa coefficients were used to study the degree of inter-annotator agreement for both annotation styles (for the ordered and unordered case). Kappa coefficients are based on the idea of rating the proportion of pairs of annotators in agreement (P_o) with the expected proportion of pairs of annotators that agree by chance (P_c). Thus obtaining a proportion of the agreement actually achieved beyond chance ($P_o - P_c$) with all possible agreements that are not by chance ($1 - P_c$):

$$\kappa = \frac{P_o - P_c}{1 - P_c} \tag{3.1}$$

For the Thesis four different Kappa coefficients were used (Figure 3.2) with which two main issues have been studied: i) the impact of annotator bias, that is, given a fixed number of agreements, the effect that the distribution of disagreements between categories has in the Kappa value; and ii) the level of importance of all possible disagreements in our task, i.e. disagreement between emotions which are easily distinguishable should have a more negative impact in the Kappa coefficient than disagreements in very different categories.

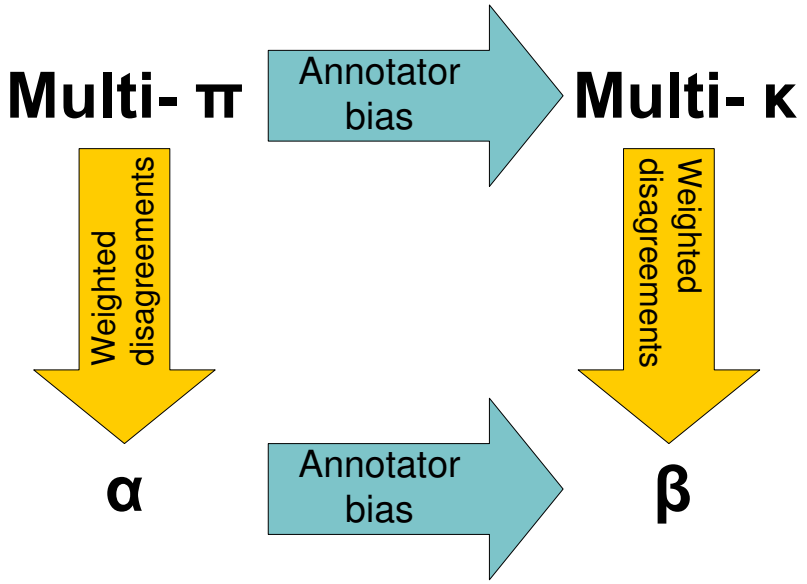


Figure 3.2. *Kappa coefficients used in the experiments*

The simplest Kappa coefficient used was proposed by Fleiss (1971), as a generalization for multiple annotators of the two-coders Scott's π (Scott, 1955). There has been much confusion in the literature about Fleiss' Kappa, as many authors have reported it as a generalization of Cohen's κ (Cohen, 1960) instead. This is further discussed by Artstein and Poesio (2005), who made a considerable effort to clarify the definitions of the different Kappa coefficients. In order to avoid inconsistencies their notation will be followed for all the Kappa coefficients employed in the chapter. In particular, Fleiss' Kappa has been noted as multi-π.

The calculation of multi-π is based on Equation 3.1, where the observed agreement (P_o) is computed as the number of cases in which two different annotators agreed to annotate a particular utterance with the same emotion:

$$P_o = \frac{1}{UA(A-1)} \sum_{u=1}^{U} \sum_{e=1}^{E} n_{ue}(n_{ue}-1) \qquad (3.2)$$

In Equation 3.2, U is the number of utterances to be annotated, A the number of annotators, E the number of emotions, and n_{ue} the number of times the utterance 'u' was annotated with emotion 'e'.

Fleiss assumes that all annotators share the same probability distribution. This means that the probability that an annotator classifies an utterance 'u' with a particular emotion 'e', can be computed as the overall probability of annotating 'u' as 'e'. This global probability was computed as the total number of assignments to emotion 'e' made by all annotators (n_e in Equation 3.3) divided by the total number of assignments ($U \cdot A$). Chance agreement (Equation 3.3) was then computed as the probability that any pair of coders annotated the same utterance with the same emotion, which was assumed to be the joint probability of each of them making such assignment independently, as annotators judged all utterances independently from each others.

$$P_c^{\pi} = \sum_{e=1}^{E} \left(\frac{1}{UA} n_e \right)^2 \tag{3.3}$$

The calculation of multi-π assumes that each annotator follows the same overall distribution of utterances into emotions. However, such a simplification may not be plausible in all domains due to the effect of the so-called *annotator bias* in the Kappa value. In our experiments, the annotator bias can be defined as the extent to which annotators disagree on the proportion of emotions, given a particular number of agreements. With the rest of the parameters fixed, the Kappa value increases as the bias value gets higher, that is, when disagreement proportions are not equal for all emotions and there is a high skew among them. This is the so-called *Kappa second paradox*. Different studies of its impact can be found in the literature, e.g. Feinstein and Cicchetti (1990), Cicchetti and Feinstein (1990), Lantz and Nebenzahl (1996), and Artstein and Poesio (2005).

To study whether inclusion of the different annotating behaviours could improve the Kappa values, Davies and Fleiss (1982) Kappa was calculated, which has been noted as multi-κ, following the study of Artstein and Poesio (2005). As happens with multi-π, the calculation of multi-κ also relies on Equation 3.1, and has the same observed agreement (Equation 3.2). However, for the chance agreement, it includes a separate distribution for each annotator. Thus, in this case the probability that an annotator 'a' classifies an utterance 'u' with emotion 'e' is computed with the observed number of utterances assigned to emotion 'e' by that annotator (n_{ae}), divided by the total number of utterances (U). The probability that two annotators agree

in annotating an utterance 'u' with emotion 'e' is again the joint probability of each annotator doing the annotation independently:

$$P_c^{\kappa} = \frac{1}{\binom{A}{2}} \sum_{e=1}^{E} \sum_{j=1}^{A-1} \sum_{k=j+1}^{A} \frac{n_{a_j e}}{U} \frac{n_{a_k e}}{U} \tag{3.4}$$

Despite of including differences between annotators, multi-κ gives all disagreements the same importance. In practice, all disagreements are not equally probable and do not have the same impact on the quality of the annotation results. For example, in our experiments, a disagreement between *neutral* and *angry* is stronger than between *neutral* and *doubtful*, because the first two categories are more easily distinguishable.

To take all this information into account weighted Kappa coefficients have been used (Cohen, 1968; Fleiss and Cohen, 1973), which put the emphasis on disagreements instead of agreements[2]. Their calculation is based on Equation 3.5 (equivalent to Equation 3.1):

$$\kappa_w = 1 - \frac{\overline{P}_o}{\overline{P}_c} \tag{3.5}$$

where $\overline{P}_o$ indicates observed disagreement, and $\overline{P}_c$ disagreement by chance. For all the coefficients used, the observed disagreement has been calculated as the number of times each utterance 'u' was annotated with two different emotions e_j and e_k by every pair of annotators, weighted by the distance between the emotions:

$$\overline{P}_o = \frac{1}{UA(A-1)} \sum_{u=1}^{U} \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} n_{ue_j} n_{ue_k} distance(e_j, e_k) \tag{3.6}$$

Consequently, the computation of the weighted coefficients implies employing distance metrics between the four emotions used for annotation (*neutral*, *angry*, *bored* and *doubtful*). To do so, the discrete list of emotions have been arranged within a continuous space, using the bidimensional activation-evaluation space (Russell, 1980). Emotions form a circular pattern

[2]An alternative calculation based on agreements can be found in (Sim and Wright, 2005)

in this space. This is why other authors proposed a representation based on angles and distance to the centre. Taking advantage of this circular disposition, angular distances between the emotions studied have been used for the calculation of the weighted Kappa coefficients. Instead of establishing our own placement of the emotions in the space, some already established angular dispositions were employed to avoid introducing measurement errors. With this purpose, the list of 40 emotions with their respective angles proposed by Plutchik (1980) was used, which has been widely accepted and used by the scientific community. In this list, *bored* (136.0º) and *angry* (212.0º) were explicitly contemplated, but this was not the case for *doubtful*. The most similar emotions found were "uncertain", "bewildered" and "confused", which only differentiated in 2º in the circle. "Uncertain" (139.3º) was chosen because it was the one that better reflected the emotion wanted to be annotated. However, other authors like Scherer (2005) have explicitly considered *doubtful* as an emotional state. Plutchik (1980) did not reflect neutral in his list as it really is not an emotion but the absence of emotion. Instead, he used a state called "accepting" as the starting point of the circle (0º), which was used as *neutral* in our experiments.

The distance between the four emotions was calculated in degrees with the angle that each of them formed in the circle. The smallest angle between the emotions being considered (x or 360-x) was always chosen. This way, the distance between every two angles was always between 0 and 180 degrees. For the calculation of the Kappa coefficients, distances were converted into weights with values between 0 and 1. A 0 weight (which corresponds to 0º distance in the proposed approach) implies annotating the same emotion, and thus having no disagreement. On the contrary, weight=1 (180º distance) corresponds to completely opposite annotations and thus maximum disagreement. The resulting distances and weights are listed in Table 3.2.

Angle/ Weight	Neutral	Angry	Bored	Doubtful
Neutral	0.00º / 0.00	148.00º / 0.82	136.00º / 0.75	139.30º / 0.77
Angry	148.00º / 0.82	0.00º / 0.00	76.00º / 0.42	72.70º / 0.40
Bored	136.00º / 0.75	76.00º / 0.42	0.00º / 0.00	3.30º / 0.02
Doubtful	139.30º / 0.77	72.70º / 0.40	3.30º / 0.02	0º / 0.00

Table 3.2. *Distance between the emotions considered*

There is not a consensus in the scientific community about the properties of the distance measures. However, Artstein and Poesio (2005) propose some constraints: the distance between a category and itself should be minimal and the distance between two categories should not depend on the order (i.e. distance from A to B should be equal to distance from B to A). As can be observed by the symmetry of the table, our distance measures and weights follow these restrictions:

- The angle an emotion forms with itself is 0º

$$\forall e \in E, distance(e, e) = 0$$

- The angle between emotion A and emotion B is the same in both directions (as it was established to choose the minimal angle):

$$\forall e_A, e_B \in E, distance(e_A, e_B) = distance(e_B, e_A)$$

As can be observed in Table 3.2, the highest distances were between non-neutrals and neutral. Thus, when calculating weighted Kappa coefficients, disagreements in which an annotator judged an utterance as neutral and the other as non-neutral were given more importance than for example an *angry* vs. *bored* disagreement.

Two weighted Kappa coefficients were calculated: Krippendorff's α (Krippendorff, 2003) and Artstein and Poesio's β. Both of them shared the same observed disagreement calculation (Equation 3.5), for α disagreement by chance was:

$$\overline{P}_c^{\alpha} = \frac{1}{UA(UA-1)} \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} n_{e_j} n_{e_k} distance(e_j, e_k) \qquad (3.7)$$

As can be observed in Equation 3.7, this coefficient does not consider annotator bias. This was solved for the β coefficient, with which also the

observed behaviour of each annotator was measured:

$$\overline{P}_c^{\beta} = \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} \left[\frac{1}{U^2 \binom{A}{2}} \sum_{m=1}^{A-1} \sum_{n=m+1}^{A} n_{a_m e_j} n_{a_n e_k} distance(e_j, e_k) \right] \quad (3.8)$$

The results for each described coefficient are listed in Table 3.3 and discussed in the next section.

Coefficient	Unordered	Ordered
multi-π	0.3256	0.3241
multi-κ	0.3355	0.3256
α	0.3382	0.3220
β	0.3393	0.3237

Table 3.3. *Values of the Kappa coefficients for unordered and ordered annotation schemes*

3.3.2. Discussion of human annotation results

As previously commented, one of the difficulties of emotion recognition in spoken dialogue systems is that in most application domains the corpora obtained are very unbalanced, because there is usually a higher proportion of neutral than emotional utterances (Forbes-Riley and Litman, 2004a; Morrison et al., 2007). This is in accordance with our experimental results since, on average among the nine annotators, more than 85.00% of utterances were annotated as *neutral*. It was also observed that this proportion is affected in 3.40% of the cases by the annotation style. Concretely, for the ordered annotation, 87.28% were tagged as *neutral*, whereas for the unordered annotation the corpus was even more unbalanced: 90.68% of the utterances were annotated as *neutral*. Figure 3.3 shows the proportion of non-neutral emotions tagged by the 9 annotators. As can be observed, the ordered annotation style yielded a greater percentage for the *bored* category: 39.00% more than in the unordered style. The figure also shows that the *angry* category is substantially affected by the annotation style (i.e. ordered vs. unordered): 70.58%

more *angry* annotations were found in the ordered annotation style. On the contrary, the *doubtful* category is virtually independent of the annotation style: only 2.75% more doubts were found in the unordered annotation.

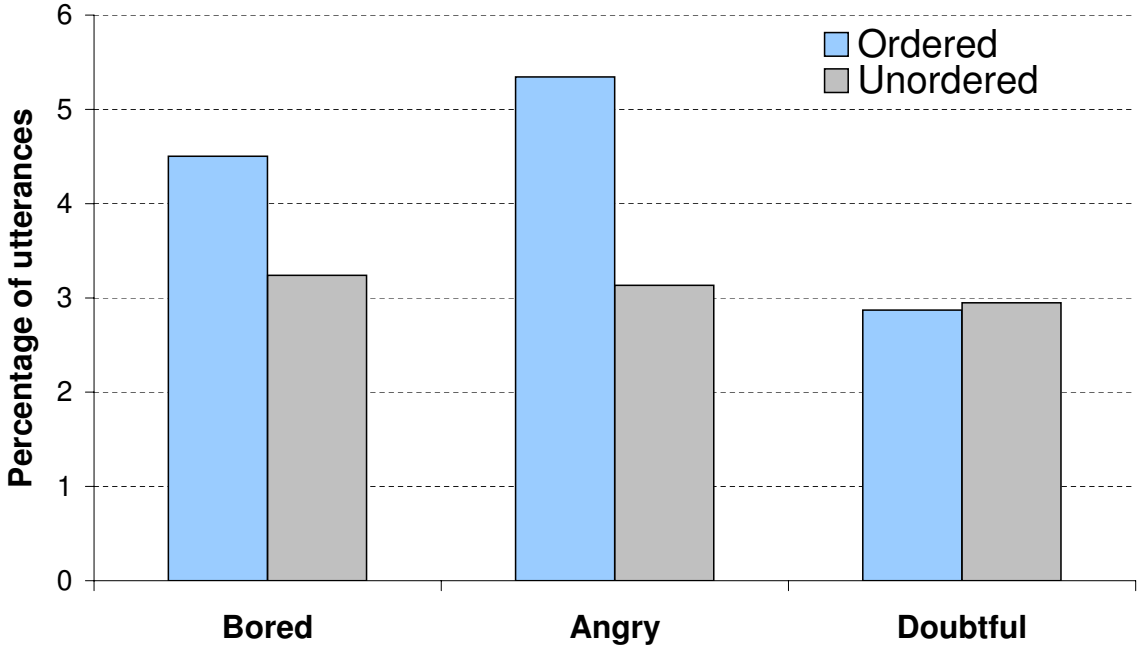


Figure 3.3. *Proportion of non-neutral annotated utterances*

A plausible reason for these results is that the incorporation of context in the ordered case influences the annotators in assigning the utterances belonging to the same dialogues into the same emotional categories. This way, there are no very noticeable transitions between consecutive utterances. For example, if anger is detected in one utterance then the next one is probably also annotated as *angry*. Besides, the context allows the annotators to have information about user's speaking style and the interaction history. In contrast, in the unordered case the annotators only have information about the current utterance. Hence, sometimes they cannot tell whether the user is either angry or he normally speaks loudly and fast.

Thus, it is an important fact to be taken into account when annotation is carried out by non-expert annotators, which is the most common, cheapest and least time consuming method. In addition, when listening to the corpus in order, the annotators had information about the position of the current user turn within the whole dialogue, which also gives a reliable clue to the user's state. For example, a user is more likely to get bored after a long

dialogue, or to become angry after many confirmation requests.

As can be observed in Table 3.3, the values of the different Kappa coefficients also vary slightly depending on the annotating scheme used. In the unordered case, both taking into account annotator bias (multi-κ vs. multi-π, and β vs. α), and weighting disagreements (β and α vs. multi-κ) improves the agreement values. However, in the ordered case only taking into account annotator bias enhances the agreement values, whereas weighting the disagreements reduces Kappa. This is a consequence of the increment of non-neutral annotations already discussed. Taking into account that the great majority of agreements occur when annotators tag the same utterance as neutral (as can be observed in Figure 3.4), an increment in the number of emotions annotated as non-neutral provokes more discrepancies among the annotators and thus reduces the Kappa value.

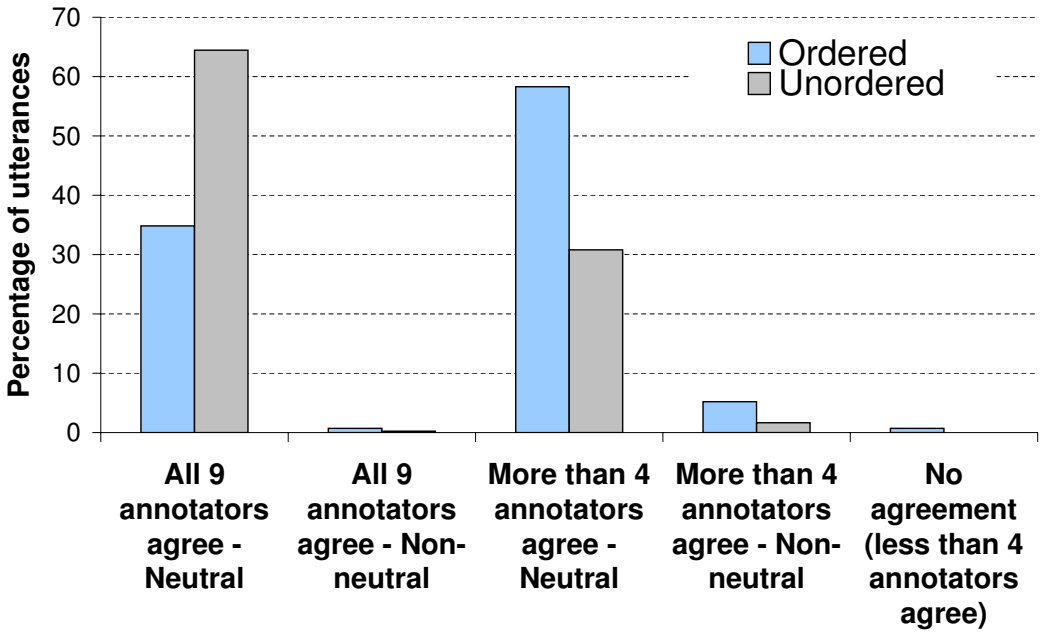


Figure 3.4. *Percentage of utterances in which annotators agree*

Furthermore, as can be observed in Figure 3.5 most of the disagreements occur between neutral and non-neutral categories, which are the emotions with higher distances according to our weighting scheme (Table 3.2), thus provoking weighted agreements to be lower in the case of the ordered scheme.

To study the effect of annotator bias, pair wise agreement between all annotators was measured. As can be observed in Figure 3.5 there were no annotators who had a significantly poor agreement with the rest. However, when the annotation results were examined, it was found that there were remarkable differences between those annotators who were used to the Andalusian dialect[3] (in which the utterances were pronounced) and those who were not so accustomed. As previously explained, the corpus was recorded from user interactions with the UAH system. The users were mainly students and professors at the University of Granada, which is in south eastern Spain. The way these users express themselves is influenced by the Eastern Andalusian dialect (Gerfen, 2002; O'Neill, 2005), which although similar to Spanish Castilian has several differences such as a faster rhythm and a lower expiratory strength. In our group of annotators, 6 were used to the Andalusian dialect (annotators 1, 2, 3, 4, 6 and 9 in Figure 3.5) and 3 were not (annotators 5, 7 and 8 in the figure).

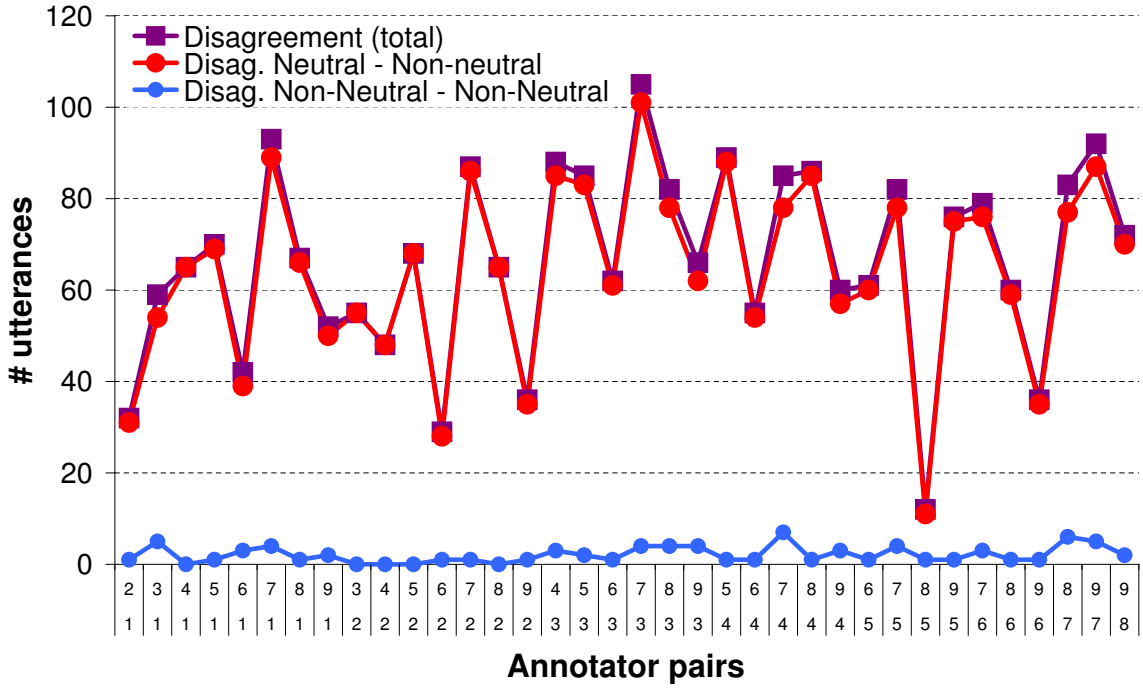


Figure 3.5. *Pair wise disagreement between annotators with the ordered scheme*

[3]Spanish spoken in Southern Spain.

Figure 3.6 shows, for the total number of annotations made in each category, which percentage corresponds to each type of annotators. As can be observed, in all the cases but one (especially in those obtained employing the ordered scheme), the annotators not used to the Andalusian dialect marked around 50% of the emotions encountered for the emotional category. This is caused by the confusion of characteristics of the dialect with emotional cues, for example confusing the Andalusian fast rhythm with an indication of anger.

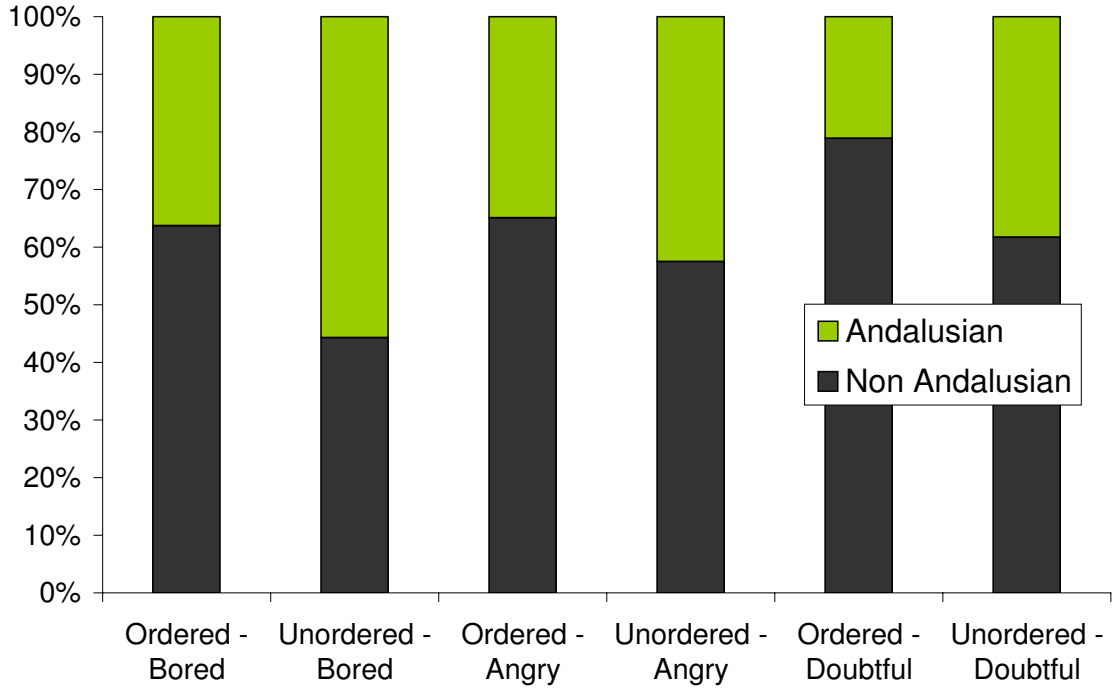


Figure 3.6. *Proportion of annotated emotions depending on dialect*

The effect on the annotation schemes for both kinds of annotators was studied and the results obtained are shown in Table 3.4. As can be observed, the annotators used to the Andalusian dialect obtained Kappa values for both annotation schemes which were more similar (ranging between 0.3234 to 0.3621). For these annotators, the Kappa values were smaller for the ordered scheme because there were fewer utterances annotated as neutral.

	Andalusian annotators		Non-andalusian annotators	
	Unordered	Ordered	Unordered	Ordered
multi-π	0.3608	0.3234	0.3734	0.5593
multi-κ	0.3621	0.3275	0.3746	0.5598
α	0.3595	0.3248	0.3644	0.5691
β	0.3607	0.3265	0.3703	0.5697

Table 3.4. *Kappa values for the different annotator types*

On the contrary, annotators not used to the Andalusian dialect had very different Kappa values depending on the annotating scheme used: in the ordered case values ranged from 0.5593 to 0.5697 whereas in the unordered these were between 0.3639 and 0.3746. This is due to a big decrement of the chance agreement. As shown in Figure 3.7, the observed agreement was more or less constant, whereas the chance agreement drastically decreased in the ordered scheme.

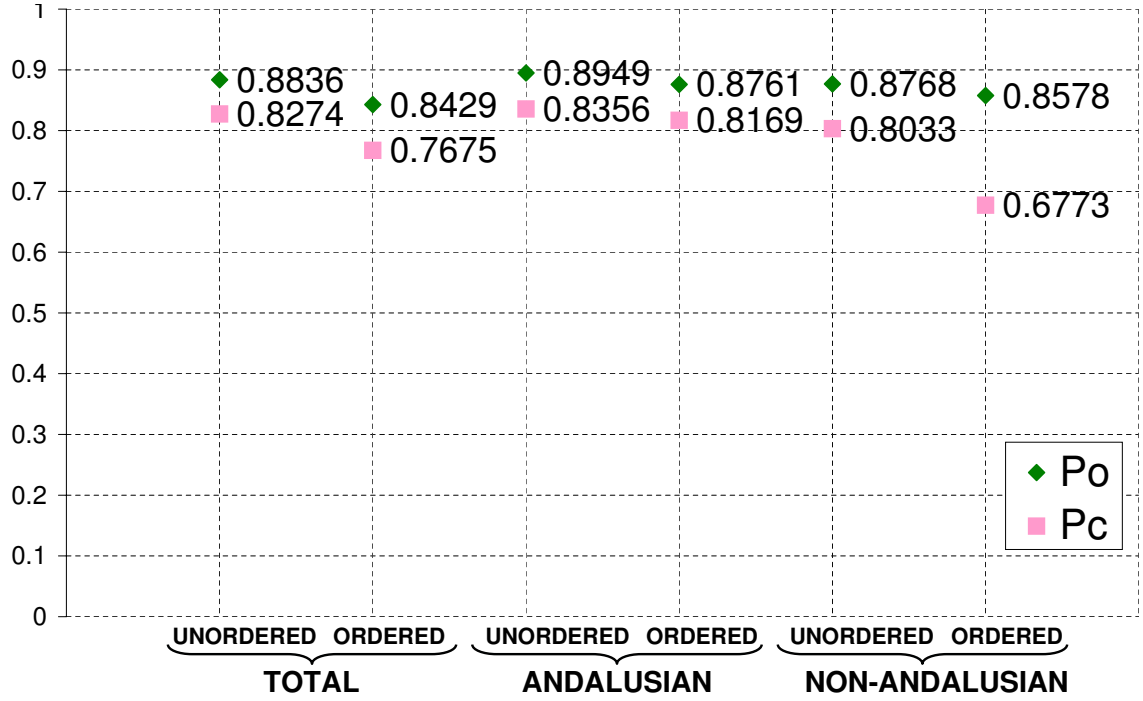


Figure 3.7. *Relative values of agremment by chance and observed agreement for multi-κ*

The most likely reason for this is the decrement in the number of neutrals annotated by annotators not used to Andalusian. This happens for both annotation schemes, but the number of neutrals annotated is higher in the unordered one, and that is why results are more similar to those obtained by Andalusian annotators with the unordered annotation scheme. Even though the number of non-neutral annotations increased proportionally with the decrement of neutrals, the unbalancement of the corpus made the probability of agreeing by chance in the neutral emotion more important in the computation of the overall agreement by chance. For example, in the case of multi-κ, agreement by chance (P_c) was calculated as the sum of agreeing by chance in each emotion ($P_c = P_c^{neutral} + P_c^{bored} + P_c^{angry} + P_c^{doubtful}$). The values for agreeing by chance when annotators not used to Andalusian used the ordered scheme were:

- $P_c^{neutral} = 0.6645$,
- $P_c^{bored} = 0.0052$,
- $P_c^{angry} = 0.0069$ and
- $P_c^{doubtful} = 0.0008$.

For the rest of annotators these values were:

- $P_c^{neutral} = 0.8137$,
- $P_c^{bored} = 0.0010$,
- $P_c^{angry} = 0.0014$ and
- $P_c^{doubtful} = 0.0008$.

Thus, $P_c^{neutral}$ was the determining factor in obtaining the global P_c.

The situation in which although having an almost identical number of agreements, the distribution of these across the different annotation categories deeply affects Kappa, is typically known as the *first Kappa paradox*. This phenomenon establishes that other things being equal, Kappa increases with more symmetrical distributions of agreement. That is, if the prevalence of a category compared to the others is very high, then the agreement by

chance (P_c) is also high and the Kappa is considerably decremented (Feinstein and Cicchetti, 1990; Cicchetti and Feinstein, 1990).

As already reported by other authors, e.g. Feinstein and Cicchetti (1990), the first Kappa paradox can drastically affect Kappa values and thus must be considered in its interpretation. There is not an unique and generally accepted interpretation of the Kappa values. One of the most widely used is the one presented by Landis and Koch (1977), which makes a correspondence between intervals for Kappa values and interpretations of agreement. Following this approach, our experimental results indicate fair agreement for both annotating schemes and with the four different Kappa coefficients. Alternatively, Krippendorff (2003) established 0.65 as a threshold for acceptability of agreement results. Hence, considering this value the 0.3393 highest Kappa obtained would not be acceptable. However, most authors seem to agree in that using a fixed benchmark of Kappa intervals does not provide enough information to make a justified interpretation of acceptability of the agreement results. In order to provide a more complete framework, some authors like Dunn (1989), propose to place Kappa into perspective by reporting *maximum*, *minimum* and *normal* values of Kappa which can be calculated from the observed agreement (P_o) as follows (Lantz and Nebenzahl, 1996):

$$kappa_{max} = \frac{P_o^2}{(1 - P_o)^2 + 1} \tag{3.9}$$

$$kappa_{min} = \frac{P_o - 1}{P_o + 1} \tag{3.10}$$

$$kappa_{nor} = 2P_o - 1 \tag{3.11}$$

The obtained Kappa values (Table 3.3) are shown in Figure 3.8 with their $kappa_{max}$, $kappa_{min}$ and $kappa_{nor}$ values, where *normal* values are marked in italics and the actual values obtained in bold.

As can be observed in the figure, for the same observed agreement, the possible values of Kappa can deeply vary from $kappa_{min}$ to $kappa_{max}$ depending on the balancement of the corpus. $Kappa_{max}$ is obtained when maximally skewing disagreements while maintaining balanced agreements, whereas $kappa_{min}$ is obtained when agreements are skewed and disagreements balanced. $Kappa_{nor}$ does not correspond to an ideal value of Kappa, but rather to symmetrical distributions of both agreements and disagree-

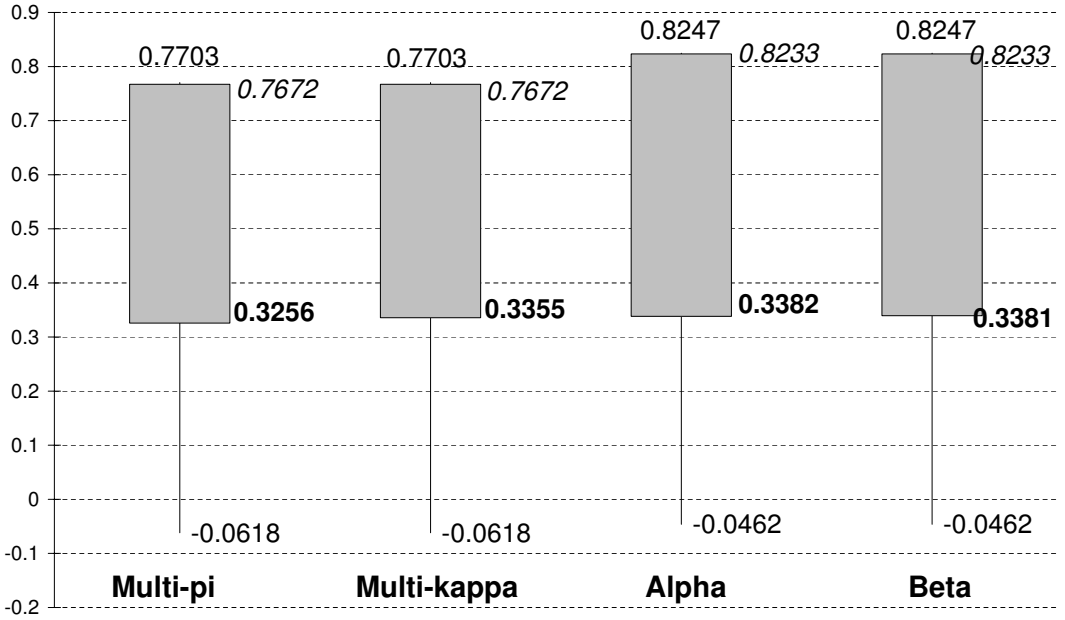
ments. It can be observed in the figure how the displacement between actual and normal values is smaller in the ordered scheme (Figure 3.8(b)). Thus, this scheme does not only allow recognizing more non-neutral emotions, but also obtaining Kappa values which, although smaller than in the unordered scheme in absolute value, are much closer to the *normal* and *maximum* agreement values attainable and further from the *minimum*.

As stated in (Lantz and Nebenzahl, 1996), departures from $kappa_{nor}$ value indicate asymmetry in agreements or disagreements depending in if they are closer to the minimum or maximum value respectively. In Figure 3.8, the shift between the observed and the normal Kappa values is represented with a box. The results corroborate that presenting Kappa values is more informative when they are put into context, as it is obtained a valuable indicative of possible unbalancements that has to be considered to reach appropriate conclusions about reliability of the annotations. For example, in our case there were significant departures from $kappa_{nor}$ in all cases, which corroborates that there was a big asymmetry in the categories. This is due to the prevalence phenomenon previously discussed (first Kappa paradox).

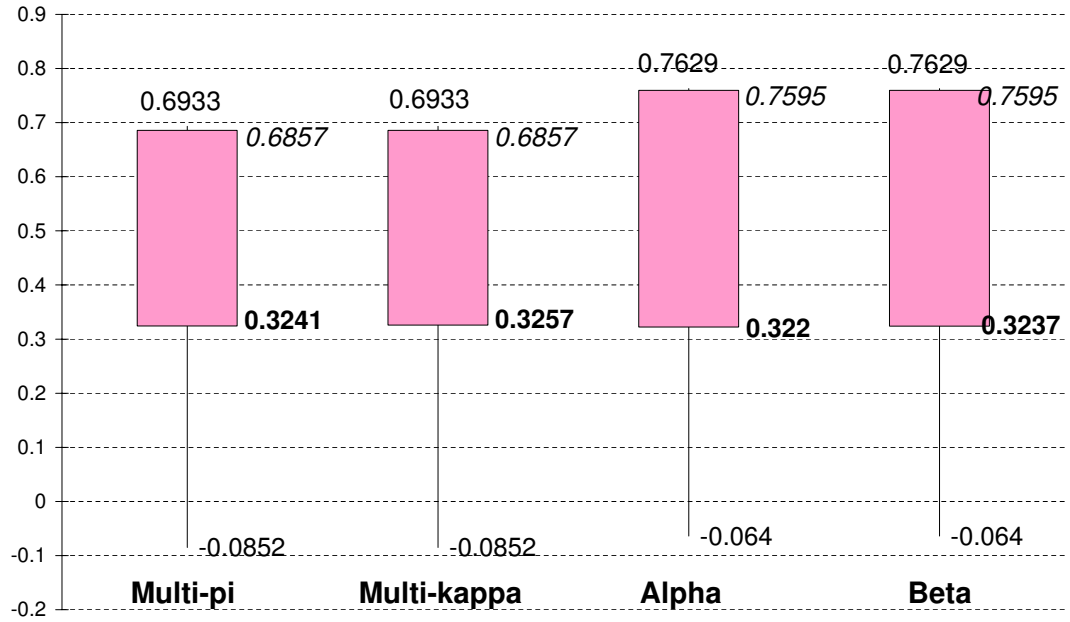
As discussed before, prevalence appeared as an unavoidable consequence of the natural unbalancement of non-acted emotional corpora, where the neutral category is clearly predominant. Thus, approaches based uniquely on already established values of acceptability (Landis and Koch, 1977; Krippendorff, 2003) are not suitable for our application domain. Some authors have already reported additional measures to complement the information provided with the Kappa coefficients. For example, Forbes-Riley and Litman (2004a) report on both observed agreement and Kappa, whereas Lee and Narayanan (2005) report on Kappa along with an hypothesis test.

Although reported Kappa values in emotion recognition employing unbalanced corpora are usually low, e.g. from 0.32 to 0.42 in Shafran et al. (2003) and below 0.48 in Ang et al. (2002) and Lee and Narayanan (2005), there is not a deep discussion about the problematic of Kappa values in the area, not even in papers explicitly devoted to challenges in emotion annotation, e.g. Devillers et al. (2005). Furthermore, even when other agreement measures are reported along with Kappa, e.g. Forbes-Riley and Litman (2004a) and Lee and Narayanan (2005), there is only one Kappa coefficient calculated (usually multi-π) and no discussion about why there is such a big difference between the Kappa values and the other measures reported. Thus,

the study presented may be one of the first reports about different Kappa values and the issues related to their use and interpretation in annotation of real emotions.



(a) Unordered scheme



(b) Ordered scheme

Figure 3.8. *Kappa maximum, minimum, normal (italics) and observed (bold) values*

Finally, to obtain a more approximate idea about the real level of agreement reached by the nine annotators, the values of observed agreement are reported in Table 3.5, which has been used along with Kappa by other authors in different areas of study (Ang et al., 2002; Forbes-Riley and Litman, 2004a). As can be observed in the table, in all the cases the observed agreement was above 0.85. This measure does not contemplate the effect of prevalence (see Figure 3.7), and thus values were not higher for the annotators not used to the Andalusian dialect in the ordered case.

Annotation scheme	User type	Observed agreement	Weighted observed agreement
Unordered	Total	0.8836	0.9117
	Andalusian	0.8950	0.9197
	Non-andalusian	0.8767	0.9050
Ordered	Total	0.8429	0.8800
	Andalusian	0.8761	0.9049
	Non-andalusian	0.8578	0.895

Table 3.5. *Observed agreement values*

From all the previous results it is possible to conclude that employing the ordered scheme allowed the annotation of more non-neutral emotions. Unfortunately, this translates in lower Kappa coefficients as most of the agreements occur for neutrals. These low Kappas indicate that multiple annotators should be used for annotating natural emotions to obtain reliable emotional corpora. One possible way to overcome the problem of high chance agreements, consists in maximizing the observed agreement. For example, Litman and Forbes-Riley (2006) propose the usage of "consensus labelling", i.e. to reach a consensus between annotators until a 100% observed agreement is obtained.

In our case, the computed Kappa values were useful to compare the two annotation schemes. As shown in Figure 3.8, although the Kappa value and observed agreement percentages are lower in the ordered case, it was found that it can be useful to obtain results which are closer to the maximum achievable. It can also be deduced from our study that evaluating the reliability of an emotion annotation process where agreements are so highly skewed, can lead to very low Kappas (Table 3.3) that are far from the high

agreement values observed (Table 3.5). Hence, it was necessary to include other sources of information like observed agreement and minimal, maximal and normal values along with the values obtained for the different Kappa coefficients in order to make meaningful interpretations. Besides, as shown in Table 3.5, giving a weight to the different disagreement types can considerably increment the observed agreement between annotators. A method to compute distances between such disagreements has been presented.

3.4 Automatic classification of the UAH corpus

As shown by the experimental results described in Section 3.3, contextual information is very important for human annotators. Therefore, in this section it is examined whether discrimination between emotions in machine-learned classification is also affected by this factor. The specific interest in this Thesis relies on distinguishing between emotions and thus only non-neutral emotions are input for the learning algorithms. The reason for doing this is not to reduce the effect of the corpus unbalancement, but to carry out a deeper study on the differences between the negative emotions considered. Thus, our main future work guideline will be to add a neutral vs. non-neutral emotion recognizer to build an emotion recognizer that copes with this natural skewness (see Chapter 6). The experiments in this section can be classified into two types: speech-related and dialogue-related. For the first group machine learning has been applied to distinguish the three emotions of interest (*angry*, *bored* and *doubtful*) and have measured the benefits of using the novel approach proposed for acoustic normalization to improve classification. For the second group knowledge about the context of the interaction has been considered in addition.

For comparison purposes, the first approach used is a baseline that always annotates user utterances with the same emotion regardless of the input. In the UAH corpus the most frequent emotion category is *angry*, therefore the baseline annotated each utterance with this label. The second algorithm used is a Multilayer Perceptron (MLP) (Rumelhart et al., 1986; Bishop, 2006). A topology with a hidden layer with $\frac{number\ of\ features\ +\ number\ of\ emotions}{2}$ nodes was used. The learning rate, which determines the speed of the search con-

vergence, was set at 0.3 to prevent it being too large (in which case it might miss minimums or oscillate abruptly) or too small (thus provoking slow convergence). To prevent the MLP from over fitting, the passes through the data (epochs) were restricted to 500. In addition, a validation threshold of 0.2 was set to determine the consecutive times that the validation set error could deteriorate before the training was stopped. To improve the performance, a momentum of 0.2 was introduced for the learning of the weights that configured the MLP. To train the MLPs and carry out the experimentation the WEKA toolkit (Witten and Frank, 2005) was employed. It is an open source collection of machine learning algorithms for data pre-processing, classification, regression, clustering and visualization.

For training and testing our classifier a 5-fold cross-validation technique has been used. Therefore, the experiments consisted of five trials where the corpus was randomly split into five approximately equal subsets (20% of the corpus each). Every trial used each of the partitions in turn for testing, and the remainder (80% of the corpus) for training, so that after the 5 trials every instance had been used exactly once for testing. Additionally, a tuning partition (20%) was extracted from each training partition in order to make the feature selection. Thus, the evaluation was carried using two phases. In the first one, the learning algorithms were trained with the 60% of training utterances and evaluated with the 20% employed for tuning. In the second phase, the complete training partition was used for training the MLP, and the test part (20%) for evaluation. For comparison purposes, this second step was carried out, on the one hand, employing all the features of the 80% training utterances, whilst on the other, employing only the features selected in the first step.

Finally, for all the experiments described in this section, the significance of the results was checked using the corrected paired t-tester available in the analysis tool of the Weka 3.5.4 Experimenter[4] (Witten and Frank, 2005), with 0.05 significance.

[4]t-statistic is calculated in Weka as:

$$t = \frac{\overline{d}}{\sqrt{\left(\frac{1}{k}+\frac{n_2}{n_1}\right)\sigma_d^2}}$$

In our case, with 5-fold cross-validation repeated 5 times: $k = 25$, $\frac{n_2}{n_1} = \frac{0.2}{0.8}$ and σ_d^2 is the variance on 25 differences.

3.4.1. Automatic classification based on standard acoustic features

For these experiments 60 features were used, which incorporate those typically used in previous studies (Devillers et al., 2005; Lee and Narayanan, 2005; Morrison et al., 2007). These are utterance-level statistics corresponding to the four groups set out in Table 3.6.

Category	Features
Fundamental frequency (F0)	Min, max, range, mean, median, standard deviation, slope, correlation coef., regression error, value at first voiced segment, value at last voiced segment
F1, F2, B1, B2	Min, max, range, mean, median value at first voiced segment, value at last voiced segment
Energy	Min, max, range, mean, median, standard deviation, slope, correlation coef., regression error, value at first voiced segment, value at last voiced segment
Rhythm	Rate, voiced duration, unvoiced duration, value at first voiced, number of unvoiced segments

Table 3.6. *Acoustic features used for classification*

The first group is comprised of pitch features. Pitch depends on the tension of the vocal folds and the sub glottal air pressure (Ververidis and Kotropoulos, 2006), and can be used to obtain information about emotions in speech. As noted by Hansen (1996), mean pitch values may be employed as significant indicators for emotional speech when compared to neutral conditions. All the pitch features in the voiced portion of speech have been computed. Specifically, the focus was on the minimum value, maximum value, mean, median, standard deviation, value in the first voiced segment, value in the last voiced segment, correlation coefficient, slope, and error of the linear regression that describes the line that fits the pitch contour. All the duration parameters (e.g. slope) were normalized by the utterance duration to obtain comparable results for all the utterances in the corpus. To extract the pitch the modified autocorrelation algorithm (Boersma, 1993) was used.

The second group is comprised of features related to the first two formant frequencies (F1 and F2) and their bandwidths (B1 and B2). Only the first two formants were used, because it has been empirically demonstrated

that adding information about a third frequency does not introduce any informative features, neither in real nor in acted emotional corpora (Morrison et al., 2007). These frequencies are a representation of the vocal tract resonances. Speakers change the configuration of the vocal tract to distinguish the phonemes that they wish to utter, thus resulting in shifts of formant frequencies. Different speaking styles produce variations of the typical positions of formants. In the particular case of emotional speech, the vocal tract is modified by the emotional state. As pointed out by Hansen (1996), in stressed or depressed states speakers do not articulate voiced sounds with the same effort as in neutral emotional states. The features that have been used for categories F1, F2, B1 and B2 are minimum value, maximum value, range, mean, median, standard deviation and value in the first and last voiced segments of each utterance.

Energy is considered in the third group of features. As stated by Ververidis and Kotropoulos (2006), this feature can be exploited for emotion recognition because it is related to the arousal level of emotions. The variation of energy of words or utterances can be used as a significant indicator for various speech styles, as the vocal effort and ratio (duration) of voiced/unvoiced parts of speech change. For example, Hansen (1996) demonstrated that loud and angry emotions significantly increase intensity, i.e. energy. For these features, only non-zero values of energy have been used, similarly as for pitch, obtaining minimum value, maximum value, mean, median, standard deviation, value in the first voiced segment, value in the last voiced segment, correlation, slope, and error of the energy linear regression.

The fourth group is comprised of rhythm features. These are based on the duration of voiced and unvoiced segments. A segment is considered to be unvoiced if its fundamental frequency is zero. The reason for this is that F0 equals the fundamental frequency of the glottal pulses, which are only generated in the presence of speech. Rhythm and duration features can be good emotion indicators, as shown by previous studies. For example, Boersma (1993) noted that the duration variance decreases for most domains under fast stress conditions.

Five rhythm features were calculated: speech rate, duration of voiced segments, duration of unvoiced segments, duration of longest voiced segment and number of unvoiced segments. All these features were normalized by the overall duration of the utterance. The speech rate has been computed as

the number of syllables normalized by the utterance duration. To compute the utterance duration the number of frames used has been multiplied by the frame step. Using the 60 acoustic features described above and the 5-fold cross-validation strategy, an emotion recognition rate of 35.42% was obtained of the MLP, whereas for the majority-class baseline it was 51.67%. The significance studies using a t-test with 0.05 significance showed that this difference is significant.

Not all the features employed for classification are necessarily very informative. Unnecessary features make the learning process slower and increase the dimensionality of the problem. Therefore, a feature selection process (Guyon and Elisseeff, 2003) has been carried out employing three methods. Firstly, a forward selection algorithm like that used by Lee and Narayanan (2005), which selects the B1 value at the last voiced segment and the maximum energy. Secondly, a genetic search method that starts with no attributes and uses a population size of 20, 20 generations, 0.6 crossover probability and 0.033 mutation probability. The selected features for this method were the following: F1 maximum, F1 median, B1 minimum, B1 range, B1 median, B1 in the last voiced segment, B2 minimum, B2 maximum, B2 median, energy maximum, energy range, and energy in the last voiced segment. Thirdly, the attributes were ranked using information gain as a ranking filter. The results employing this method were: energy maximum ranked with 0.50, B1 in the last voiced segment with 0.46, and other features were evaluated with 0. Taking into account the three approaches, the optimal subset was composed of B1 in the last voiced segment and energy maximum. When classifying with the selected features only, no improvements were obtained in the experiment, as the percentage of correctly classified utterances was 49.00%, which is worse than obtained with the baseline. However, this difference was not significant in the t-test, which indicates that the results for both the MLP and the baseline are equivalent after feature selection.

3.4.2. Automatic classification based on normalized acoustic features

To reproduce the user's speaking style information that the annotators had when they labelled the corpus in the ordered case, a new approach is proposed in which acoustic features are normalized around the neutral voice of the user.

For example, let us say that user 'A' always speaks very fast and loudly, and user 'B' always speaks in a very relaxed way. Then, some acoustic features may be the same for 'A' neutral as for 'B' angry, which would make the automatic classification fail for one of the users. This is what happened to the annotators who were not used to the Andalusian dialect (Section 3.3.2), as they were confused by the fast rhythm and loud speech of the speakers who recorded the corpus.

In order to carry out the proposed normalization the user's neutral voice features are obtained in each dialogue, and subtract them from the feature values obtained in the rest of the user's utterances. To calculate the neutral voice, it could have been used the average value of all utterances of that user in the corpus labelled as *neutral* by the annotators. However, our intention for future work is to integrate our emotion recognizer in the UAH system, so that it can be adaptive to user's emotions. It is impossible to carry out this computation in execution time as this would require to have all the user turns in the dialogue in advance. Therefore, the first utterance of the user was considered to be neutral, assuming that he is initially in a non-negative emotional state. Besides, the interest of the Thesis is only on emotions caused by the interaction with the system, assuming that the user is in a neutral emotional state when he starts the interaction with the dialogue system. This assumption is possible in domains not directly related to highly affective situations, such as bookings or information extraction, which are the typical application domains of spoken dialogue systems. For these application domains, dialogues in which the user is already in a negative emotional state are negligible.

The accuracy obtained with the Multilayer Perceptron using the normalized features was 53.17%. Thus, introducing acoustic context enables the MLP to improve over the results obtained by the baseline, but the improvement was not significant. Employing the features selected in Section 3.4.1 (B1 in the last voiced segment and energy maximum) 69.33% correctly classified utterances were obtained, which showed to be a significant improvement following the t-test. In the non-normalized case the feature selection did not yield any improvement. Thus, using normalized acoustic features yielded an improvement of 17.66% (69.33% recognition rate) over the baseline, which was also the best case in the non-normalized classification (Figure 3.9).

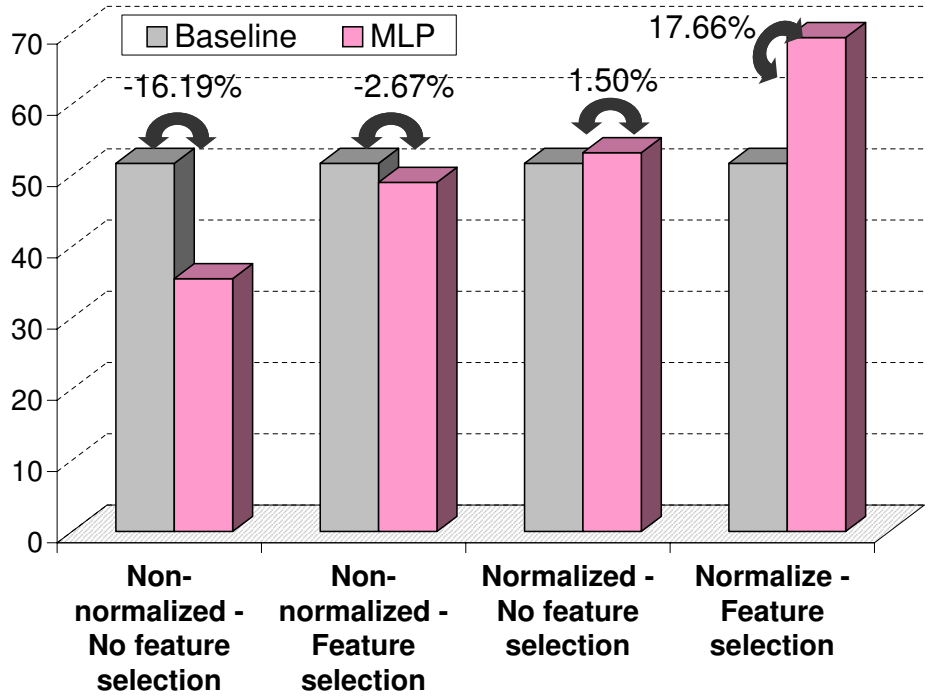


Figure 3.9. *Recognition accuracy for angry, bored and doubtful considering non-normalized vs. normalized acoustic features, and no feature selection vs. feature selection*

Thus, the normalization of traditional acoustic features yields a noticeable increase in the percentage of correctly recognized emotions with respect to the baseline. This is a very important result as, due to the natural skewness of non-acted emotional corpora, high accuracies can be obtained when directly assigning the most frequent category to all the prompts. In our case, the baseline yielded an accuracy higher than 50.00%. Only with normalization the MLP could obtain better results than the baseline, which were improved by a 17.66% when using acoustic features selection.

A study of the confusion matrices in both annotation schemes showed that the *doubtful* category was often confused with the *angry* or *bored* categories, with confusion percentages above 20% in most cases. A similar result was obtained for human annotation given that the ordered scheme did not improve the annotation of the doubtful emotion (as can be observed in Figure 3.3). These results show that contextual information affects automatic speech recognition using these classification methods, similarly as it affects human annotation.

Thus, in order to distinguish between *doubtful* and the other negative emotions, additional sources of contextual information must be added. Our proposal is to automatically recognize the three emotions using a two-step method. In the first one, acoustic information and contextual information about the user's neutral voice are used to distinguish between *angry* and *doubtfulORbored*. In the second step, dialogue context is used to discern between *doubtful* and *bored*. In the first step, the previously described normalization procedure was used to recognize *angry* vs. *doubtfulORbored*.

To optimize the results, another feature selection was carried out, in which the optimal features are those that best discriminate between *angry* and *doubtfulORbored*. Using the same feature selection algorithms, a subset comprised of three features was obtained. These were F2 median, energy maximum and energy mean. The results obtained are shown in Figure 3.10. All of them proved to be significantly better than the baseline using the t-test, except for the first case (non-normalized and no feature selection), where the results were the same order for the MLP and the baseline. The best result for *angry* and *doubtfulORbored* was achieved with feature selection in the ordered scheme, where an 80.00% accuracy was obtained.

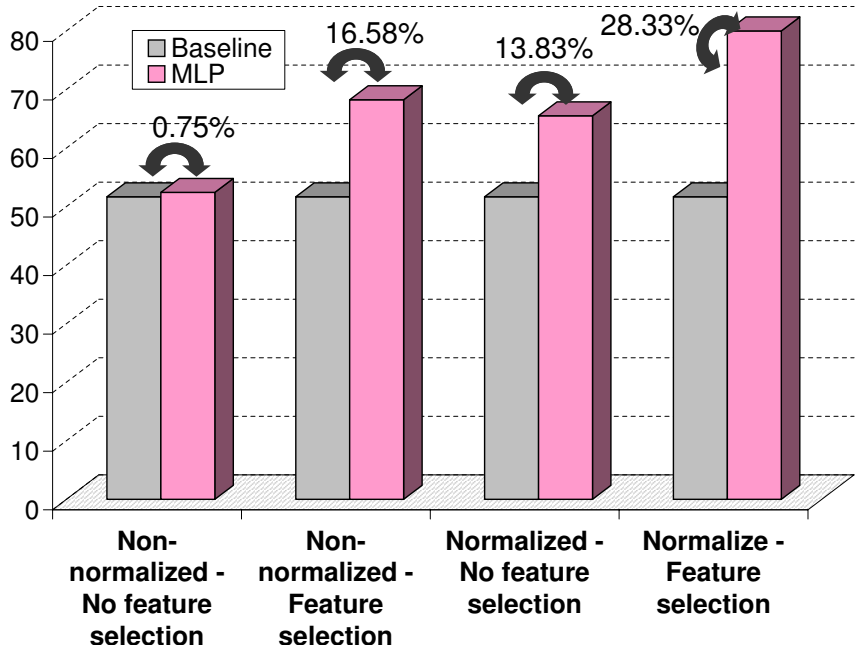


Figure 3.10. *Recognition accuracy for angry and doubtfulORbored considering non-normalized vs. normalized acoustic features, and no feature selection vs. feature selection*

With these experiments it has been shown that acoustic features normalized with neutral style of the users are preferable to the non-normalized ones, as these yield 17.66% improvement (69.33% vs. 51.67% success rate) when recognizing between the three negative emotions, as shown in Figure 3.9. Moreover, if acoustic information is used in the first recognition step which distinguishes between *angry* and *doubtfulORbored*, accuracy of 80.00% can be achieved, which represents an improvement of 28.33% over the baseline, and of 11.75% over the case when no context information about the user's neutral voice is used (Figure 3.10). In the next section, a second step is described so that dialogue context can be added to distinguish between *doubtful* and *bored*.

3.4.3. Automatic classification based on dialogue context

As discussed in Section 3.3, dialogue context was provided to human annotators by giving them the ordered sequence of utterances in each dialogue. To represent context information for automatic recognition two labels have been employed: *depth* and *width*. The first of these indicates the total number of dialogue turns, whereas the second denotes the number of additional user turns necessary to obtain a particular piece of information (e.g. a person's surname). Other authors have already studied the use of discourse structure in similar ways for other areas. For example, Rotaru and Litman (2006) studied how specific locations in the discourse structure are more prone to speech recognition errors in dialogue systems. To do so, they quantified the position of the user turns employing two similar features: "depth" and "transition". However, their approach is based on intentional models of dialogue, which consider it divided in different goals or intentions that have to be satisfied to complete a task. Thus, they define "depth" of a user's turn as the number of different intention subdialogues (or subgoals) in the previous history. This way in their approach several user turns can have the same "depth". Similarly, their "transition" feature captures the position in the discourse structure relative to the previous turn, describing it using different labels for the type of transition between the last and the current turn (e.g. if the new turn introduces a new intention or continues with the same goal).

Although also based in vertical and horizontal positions of user turns in the dialogue structure, the proposed approach employs numerical representations, which allows the automatic calculation of dialogue context from the dialogue history, without introducing new manual annotations like Rotaru and Litman's "transition" tags. Thus, a completely new calculation of the dialogue context is proposed in which the different values of the vertical component (*depth*) are not shared by different turns. Only our *width* label is concerned with the subdialogue in which the user is engaged. However, our interest is only on subdialogue types that could cause a negative emotion, which have been considered to be those in which the user can not provide any new information and has to repeat, correct or confirm data.

The computation of the dialogue context in the proposed approach is as follows:

- *Depth* is initialized to 1, given that 0 would mean that the user hangs up the telephone before he says anything. *Depth* is incremented by 1 for each new user turn and each time the interaction goes backwards (e.g. to the main menu).

- *Width* is initialized to 0, which indicates that the user does not need to repeat any information to the system initially. It is incremented by 1 for each user turn generated to confirm, repeat data or ask the system for help. It is also incremented by 1 if the user has to disambiguate his own input, even if providing new information. Then *width* is restablished to 0 when the user introduces a new piece of information.

For illustration purposes, a sample dialogue from the UAH corpus is shown below (translated from Spanish to English), where S = system, U = user:

S1> Welcome to the UAH system.
U1> I need information about professors.
S2> Please say the name of the professor.
U2> López.
S3> I have found two professors named López: Pedro López and Sara López. Who do you mean?
U3> Sara López.
S4> What information do you need about Prof. Sara López?
U4> I want her telephone number.

S5> Sorry, I did not understand. What information do you need?
U5> Telephone.
S6>The telephone number of Prof. Sara López is 958123456. Do you need any other information about professors?
U6> No thanks.
By applying the annotation scheme to this sample dialogue the following values have been obtained for *Depth* and *Width*:

U1> Depth=1, Width=0
U2> Depth=2, Width=0
U3> Depth=3, Width=1
U4> Depth=4, Width=0
U5> Depth=5, Width=1
U6> Depth=6, Width=0

It can be observed in this example that the user needed to employ two turns (U2 and U3) to make the system understand the professors' name. In turn U5 he rephrased what he said in turn U4, which solved the system misunderstanding. This is the reason why *width* is 1 for these two user turns.

This scheme is implemented automatically in the system using the dialogue history, which is stored in log files. The *depth* and *width* values of a user turn are computed checking the type of the previous system prompt. For example, as shown in Figure 3.11, *width* would only be 0 after a system prompt of type "subject_name" if the current prompt type were "subject_information". If the current prompt type were "subject_disambiguation", *width* would be incremented by 1 because an additional user turn would be needed to provide the desired subject to the system.

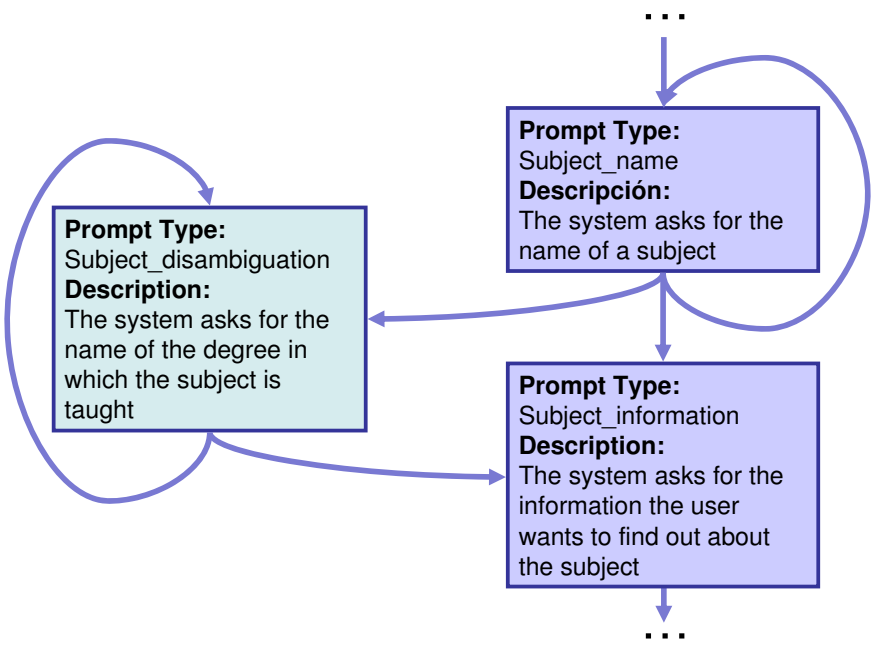


Figure 3.11. *Example of transitions between system prompts in the UAH system*

An exhaustive study of the UAH corpus showed that different users react with different emotions to the same dialogue state, in a less predictable way than initially expected. Employing one threshold for *depth* and another for *width* to distinguish independently between emotions was found to be inefficient as emotions are influenced by a mixture of the two. Furthermore, the study of the UAH corpus revealed that it is not sufficient to compute *width* considering only the previous turn or current subdialogue, but it is necessary to take into account the whole dialogue history. This differs from the results reported by Rotaru and Litman (2006), which annotate their horizontal variable ("transition") only with information about the previous system prompt. For example, in the following dialogue:

(...)
S1> Please say the name of the professor.
U1> Martín.
S2> Sorry, I did not understand. Please repeat the name.
U2> Luis Martín.
S3> Did you say Luis Marín?

U3> No, Luis Martín.
S4> What information do you need about Prof. Martín?
U4> His email address.
(...)

width would be 0 in U1, 1 in U2, 2 in U3 and 0 again in U4 because the dialogue starts to deal with a new piece of information. A high *width* value in U2 indicates a higher probability of the user being angry because of the misunderstandings and repetitions needed to make the system understand the name of the professor referred to. However, in turn U4 the user may still be angry but it has *width*=0.

This is why it has been defined another measure called *accumulated width*. Whereas *width* is a measure of the extra turns necessary to obtain a particular piece of information, *accumulated width* denotes the total number of extra turns employed in the whole dialogue up to the current utterance. *Accumulated width* is initialized to 0 and it is increased by 1 each time *width* is incremented. Thus, in the previous example, in U3 $width = 2$, which indicates that it was necessary to repeat or confirm the information of the professor's name twice. Note that in turn U4 $width = 0$ again because the system is gathering different data, namely the type of information about the professor that the user wants. Hence, accumulated width is more representative than width because it takes into account all the problematic points in the previous dialogue. For example, in U4 *accumulated width* $= 2$, which lets us know that there were 2 problematic turns before the current prompt.

The algorithm employed to classify the emotions based on the dialogue context information is as follows:

if Any of the 2 previous turns has been tagged as $ANGRY$ **then**
  $ANGRY$
else if $(D \leq 4)$ AND $(A \leq 1)$ **then**
  $DOUBTFUL$
else if $(\frac{A}{D} \geq 0.5)$ OR $((D > 4)$ AND $(\frac{A}{D} < 0.2))$ **then**
  $BORED$
else
  $ANGRY$
end if

Where 'D' denotes *depth* and 'A' the *accumulated width*. In the proposed approach, the user utterances are considered as *doubtful* when the dialogues are short and have no more than one error, as in the first stages of the dialogue is more probable that the users are unsure about how to interact with the system. An utterance is recognized as *bored* when most of the dialogue has been employed in repeating or confirming information to the system. The user can also be *bored* when the number of errors is low but the dialogue has been long. Finally, an utterance is recognized as *angry* when the user was considered to be angry in at least one of the two previous turns in the dialogue (as described at the beginning of Section 3.3.2 with human annotation), or the utterance is not in any of the previous situations (i.e. the percentage of the full dialogue length comprised by the confirmations and/or repetitions is between 20% and 50%).

When considering a baseline that always classifies utterances with the most frequent emotion, which in our case is *angry* (same baseline as in Section 3.4.2), 51.67% accuracy is obtained in distinguishing between the three emotions. This rate is improved by 13.61% employing the proposed algorithm, which attains an accuracy of 65.28%.

3.4.4. Automatic classification based on normalized acoustic features and dialogue context (two-steps method)

A method in two steps is proposed, which integrates both contextual sources: the users' neutral speaking style (Section 3.4.2) and the dialogue context (Section 3.4.3). The acoustic features normalized with the users' neutral speaking style are used to discriminate whether each utterance is *angry* or *doubtfulORbored*. Then, if an utterance is classified as *doubtfulORbored*, dialogue context information is used to distinguish between *doubtful* and *bored*. Additionally, dialogue context is used to classify utterances as *angry* if they were misrecognized in the first step. The results obtained by the two-step method are shown in Figure 3.12, and proved to be significant following the t-test.

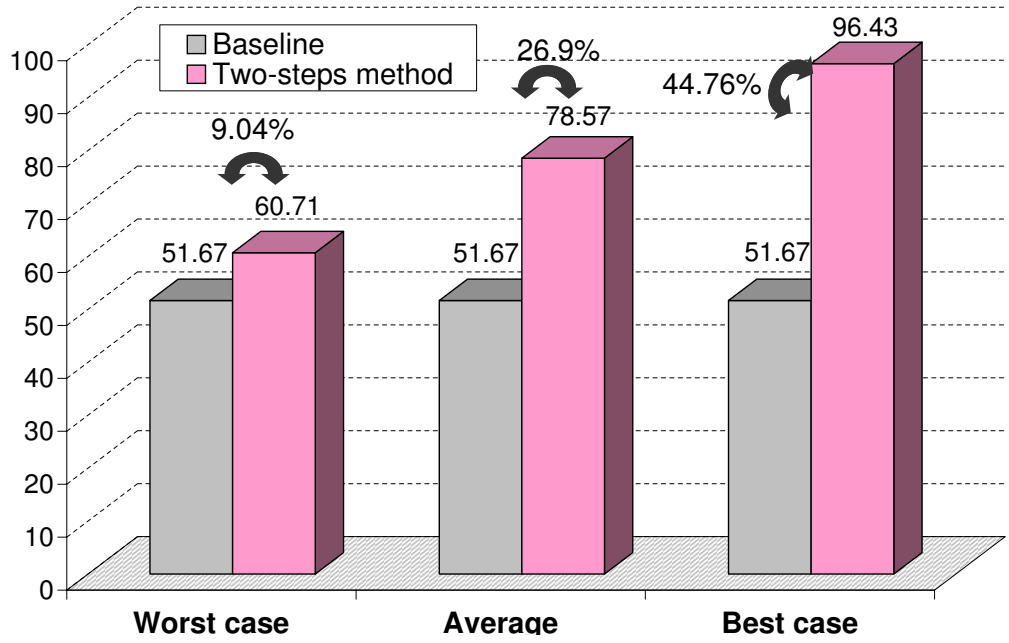


Figure 3.12. *Emotion recognition accuracy using both acoustic and dialogue context information*

Obviously, the result of the two-step method depends on the result of the first one, given that the *angry* vs. *doubtfulORbored* step can fail in the distinction of the two categories and the second step may have to categorize as *doubtful* or *bored* an utterance that belongs to the *angry* category. In the worst case, the first step can fail to recognize all the *angry* utterances, so that all the utterances are recognized as *doubtfulORbored* and passed to the second step. In this case, the recognition accuracy is 60.71%, as a mechanism to detect possible angry utterances has been incorporated to the second step (see Section 3.4.3). In an ideal best case, the first step would have 100% accuracy, and thus would correctly classify all the utterances as *angry* or *doubtfulORbored*. Thus, the second step would only have to classify the *doubtful* and *bored* utterances. The recognition rate in this case is 96.46%. However, as it was discussed in Section 3.4.2, with the UAH corpus the first step obtains a maximum 80.00% accuracy, which means that 20.00% of the *angry* utterances may be misrecognized. Employing the two-step method the recognition rate was again 96.43%. Thus, the misrecognized *angry* utterances could be correctly classified in the second step, obtaining a recognition rate for our best case in practice, which is identical to the ideal best case.

On average between the worst and best case, the two-step method obtains a 78.57% accuracy (as observed in Figure 3.13), which outperforms the baseline by 26.90%. The improvement over the baseline is 44.76% in the best case, i.e. when the first step does not fail. The average improvement over the recognition based only on neutral acoustic context is 9.24% (27.10% in the best case). If the recognition is based on dialogue context only, the average improvement is 14.29% (32.15% in the best case), and if it is based on the traditional approaches considering non-normalized acoustic features, the average improvement is 29.57% (47.43% in the best case).

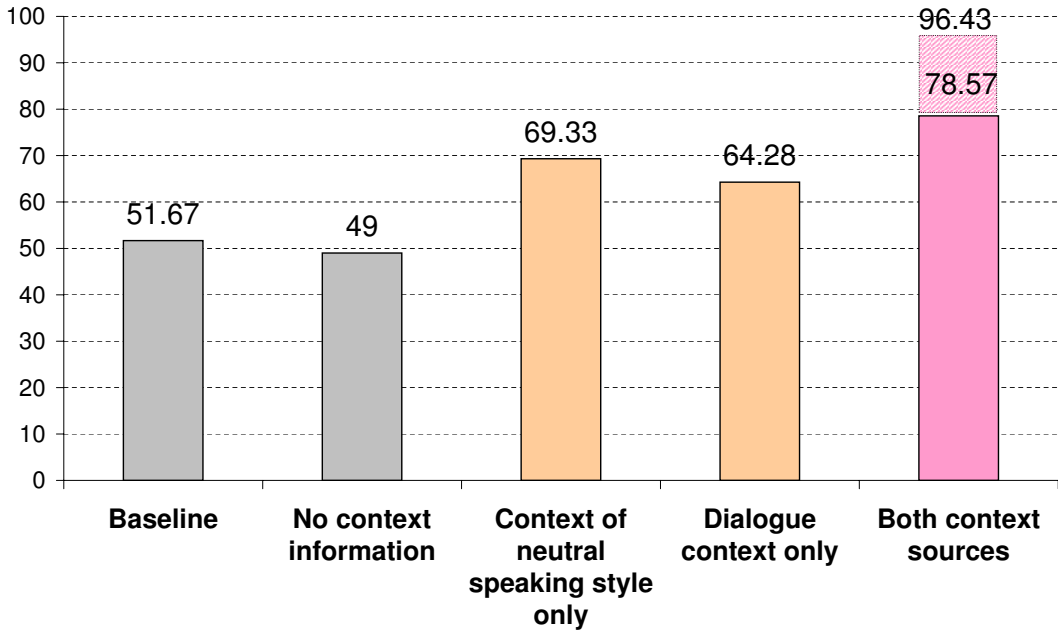


Figure 3.13. *Comparison of the recognition accuracies of the methods for automatic emotion recognition*

Thus, using only one context source (neutral voice or dialogue context), improves over both the baseline and the traditional approach where no context information is used. Besides, combining the two context sources in the proposed two-step method considerably outperforms the baseline, the traditional approach based on acoustic features without additional context sources, and the approach considering only one context source either the neutral voice of the user or the dialogue context.

3.5 Previous version of the two-steps method

Before finding the optimal approach, an detailed study was carried out of the different possibilities to recognize bored, angry and doubtful emotions in a two-step method using the proposed context sources. The main objectives were to maximize: i) the significance of the obtained results, ii) the difference between the baseline and the proposed methods in each step, and iii) the use of contextual knowledge in the whole process.

A previous version of the two-steps method distinguished firstly between *doubtful* and *angryORbored* using dialogue context, and secondly between *angry* and *bored* using the neutral speaking style of the user, as shown in Figure 3.14.

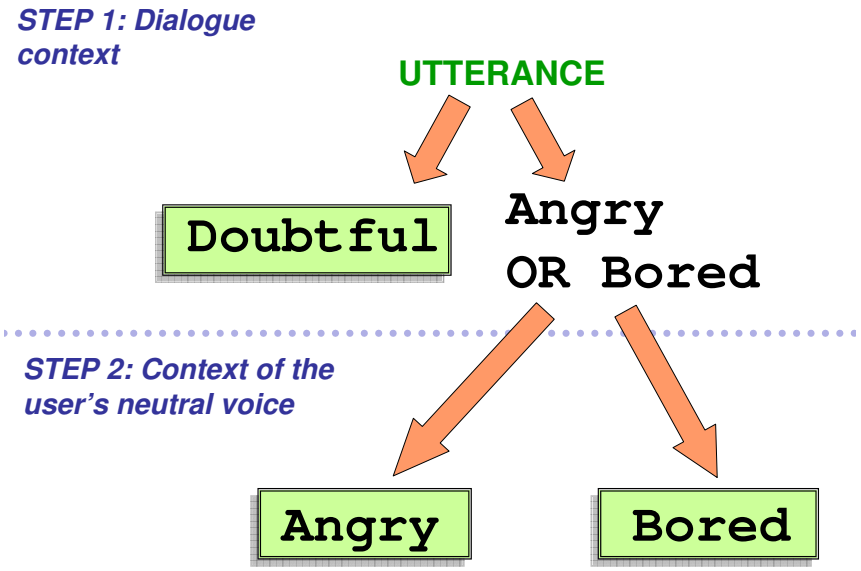


Figure 3.14. *First version of the two-steps method for automatic emotion recognition*

In this first version of the two-steps method, a static threshold T was used for dialogue context, which was computed as follows:

$$T = D + A$$

where 'D' denoted *depth* and 'A' the *accumulated width*. In this approach, a value of T greater than or equal to the threshold indicated *angryORbored*, whereas a smaller value indicated *doubtful*. Several values for the threshold were studied, which classification results are shown in Figure 3.15.

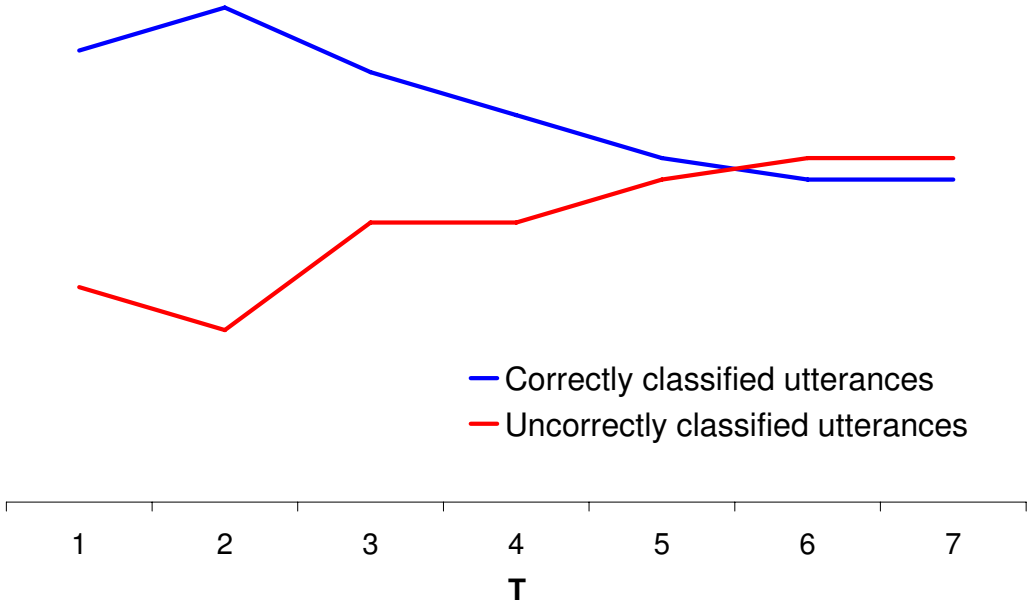


Figure 3.15. *Impact of the value of dialogue context thresholds in emotion classification success*

As can be observed, for thresholds greater than five the percentage of correctly classified utterances is smaller than the number of incorrectly classified. For lower threshold values there were more correctly than incorrectly classified utterances. However, although very low T values like T=2 yielded a higher difference between correctly and incorrectly classified instances, they were not optimal because the corpus was unbalanced (there were more utterances labelled as *angryORbored* than as *doubtful*) and thus the best results were obtained when almost all values were classified as *angryORbored*. In fact, the confusion matrices obtained showed that for values of T smaller than four, *doubtful* utterances were mostly incorrectly classified. As a result, $T = 4$ was employed as the optimal value for the threshold. Thus, the classification approach consisted in assigning *doubtful* to the user turns with $T < 4$ and *angryORbored* when $T \geq 4$; which yielded 70% classification accuracy.

Once an utterance was classified as *angryORbored*, the normalized acoustic features enabled the distinction between *bored* and *angry* following the same procedure as described for the final version of the two-steps method in Section 3.4.4. The classification rate with the acoustic features

was 85.71% for the distinction between *angry* and *bored*. Thus, a maximum 60% classification rate could be attained for the three emotions (*angry*, *bored* and *doubtful*), assuming that the first step was completely succesful (70% rate). This was 24.52% better than the case in which no context information was used, but it was worse than using only one of the context information sources separately, as shown in Figure 3.16

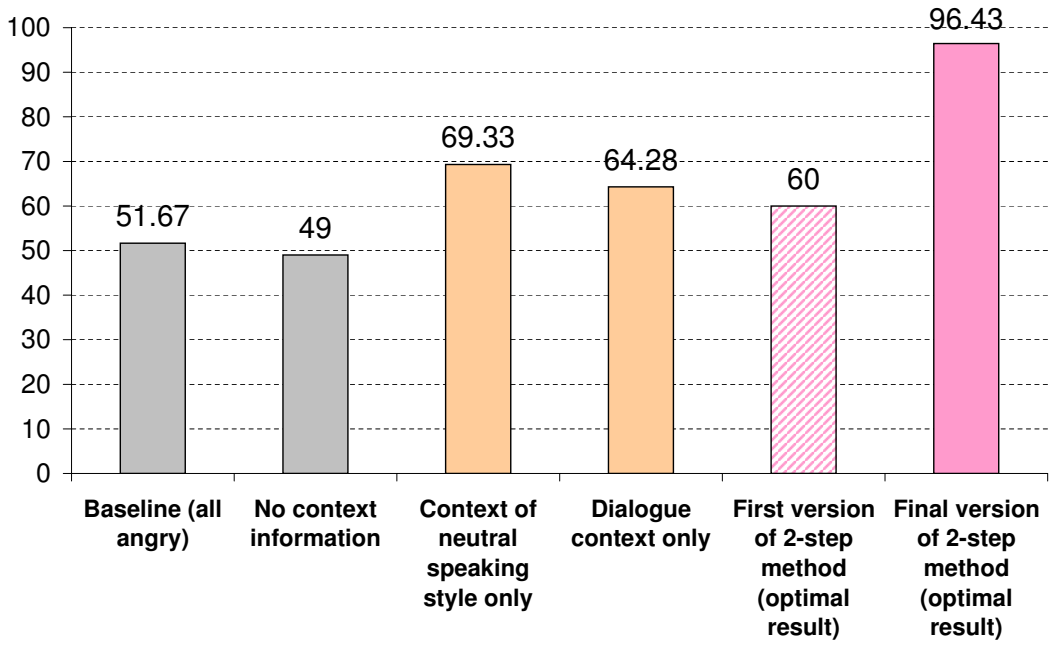


Figure 3.16. *Comparison of the accuracy of the different methods proposed for automatic emotion recognition*

After a deep study of the characteristics of the three emotions, the method was improved and the optimal version was obtained, which is the one that has been described in Section 3.4, and is shown in Figure 3.17.

In this scheme, employing acoustic information along with the context of the user's neutral voice, the utterances were classified as *angry* and *doubtfulORbored*. In a second step, those classified as *doubtfulORbored* were marked as *bored* or *doubtful* using the dialogue context. The results are totally comparable between the final two-steps method, the baseline, and each context source employed separately. This was not possible in the first version of the method where the approach used to take into account dialogue context could not distinguish between angry and bored, and thus it could not be used in isolation to distinguish between the three emotions.

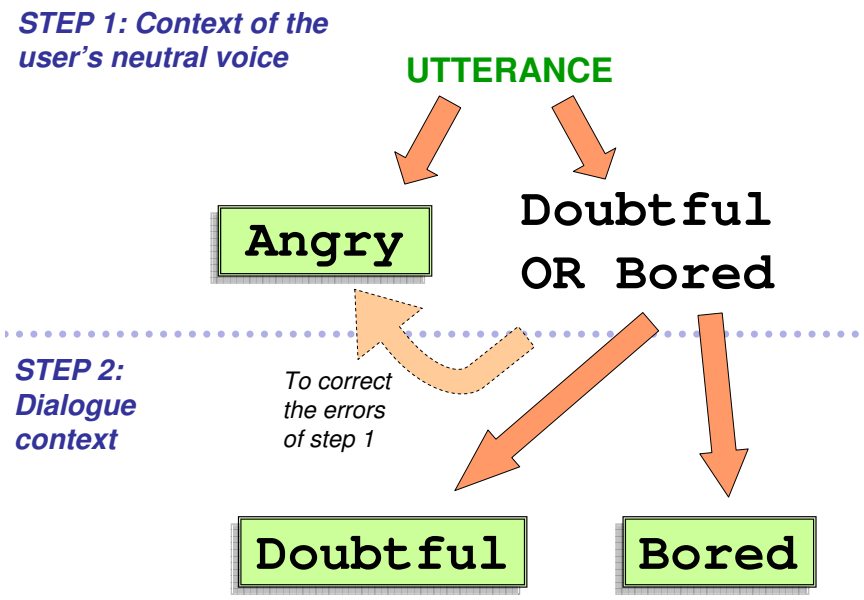


Figure 3.17. *Final version of the two-steps method for automatic emotion recognition*

3.6 Conclusions

In this chapter several experiments have been carried out to study the annotation of human emotions in a corpus collected from real interactions with the UAH dialogue system. The experiments considered both manual annotation by 9 non-expert human annotators, as well as automatic classification employing MLPs and different feature selection techniques. It was found that human annotators marked 3.40% more non-neutral emotions when they had contextual knowledge. A plausible reason for this result is that context information makes it possible to identify non-trivial emotional speech (e.g. detecting emotions expressed more subtly). On the contrary, when the traditional non-normalized acoustic features were used, only very easily distinguishable emotions were annotated. Additionally, it has been discussed the problems that the nature of non-acted emotional corpora impose in evaluating reliability of human annotations. Although deeply affected by the so-called paradoxes of Kappa coefficients, it has been studied how the inclusion of context information during annotation helps to obtain values closer to the maximum agreement rates obtainable when compared to not using any additional information.

For machine-learned classification methods, the experimental results show that, due to the natural unbalancement of the corpus, it is difficult to improve the baseline. This makes traditional recognition based on acoustic features yield results very similar to the majority-class baseline. However, as with human annotators, the emotion classification process is substantially improved when adding information about the user's neutral voice and the dialogue history. Just introducing the user's neutral acoustic context gives an improvement of 17.66%. Similarly, employing dialogue context information improved the baseline results in 12.67%. In this chapter it has been described a method in two steps to integrate both sources of contextual information. In this method, the normalized acoustic features are useful to distinguish between *angry* and *doubtfulORbored* categories with a 80.00% success rate. Once an utterance is classified as *doubtfulORbored*, the dialogue context enable us to distinguish between *doubtful* and *bored*. When the first step attains maximum accuracy, the two-step method obtains 96.43% accuracy. In the average case, the proposed method obtains 78.57% accuracy, which is 29.57% better than not using contextual information, 47.43% better in the best case (when the first step reaches its maximum performance).

In addition, the proposed methods can be employed during the running of a dialogue system as the contextual information sources can be obtained automatically and at execution time. To do so, a normalization process of the different acoustic measures with the users' neutral speaking style has been proposed; as well as a representation of the dialogue structure based on two parameters that can be numerically calculated from the information generated by the dialogue manager at run time.

Kolika jazyků znáš,
tolikrát jsi člověkem
Czech proverb

4 Cross-lingual adaptation of speech recognizers

4.1 Introduction

Cross-lingual adaptation makes it possible to employ corpora and resources already available in a language for the recognition of a different one. This allows fast and low-cost implementation of speech recognizers, although in detriment of the accuracy of the recognition result. However, this decrease in performance can be in many cases considered negligible when balanced against the cost of obtaining the resources required for building a recognizer for the new language. This approach is especially useful for minority languages or dialects in which the number of shared resources available is very limited or even not existent.

The hypothesis that we wanted to demonstrate in this chapter of the Thesis was that a fully functional system based in the Czech language can be easily and rapidly adapted for the interaction in another language without the need of building a new speech recognizer or getting involved in an arduous linguistic study, as for example morphological diacritization in (Kirchhoff and Vergyri, 2005). Thus, a new approach to reach this objective is proposed and experimental results that measure its appropriateness are presented with both a language that is similar to Czech (Slovak) and a language from a very different origin (Spanish).

Firstly, with the adaptation to the Slovak language we wanted to prove the suitability of the proposed technique to rapidly adapt an existing speech recognition system to work with a phonetically similar language. Besides, it

was also our aim to show that it is possible to take the most of the costly process of obtaining all the resources necessary to build speech recognition systems for a minority language such as Czech (spoken by approximately 12 million people), by employing it with a less-resourced language like Slovak (spoken approximately by 6 million people), and obtain high accuracy rates.

Secondly, the main contribution is the adaptation to Spanish, with which the objective was to find out whether this straightforward cross-lingual adaptation could also yield good results with languages that belong to different families and thus are phonetically less similar. As can be observed in figure 4.1[1], Czech belongs to the family of Slavic languages like Russian, concretely to the Czech-Slovak together with Slovak. On the other hand, Spanish is an Italic language like Italian or French; concretely it belongs to the West-Iberian group as Portuguese. Thus, one of the challenges of the Thesis was to obtain a satisfactory mapping for such different languages (Czech and Spanish); especially when previous researches have obtained poor results in cross-language tasks between Slavic and Italic languages. For example, this is the case for (Zgank et al., 2004), who studied Slovenian and Spanish and, based in their experimental results, recommended addressing only languages that are very similar to ensure maximal overlap of phonemes. In the literature it is frequent the usage of languages with the same roots, for example Italian and Spanish (Bonaventura et al., 1997) (Italic language family) or English and Afrikaans (Nieuwoudt and Botha, 2002) (Germanic language family).

The chapter is structured as follows. Section 4.2 presents related work in the area of cross-linguality. It describes our proposal and compares it with the state-of-the-art methods. Section 4.3 describes the previously created Czech speech recognizer which was used to recognize Slovak and Spanish, and the MyVoice system. In Section 4.4, the cross-lingual adaptation is explained for every language used in the experiments. Section 4.5 describes the experiments carried out to test the performance of the proposed technique, whereas Section 4.6 discusses the results obtained, and Section 4.7 describes the conclusions reached.

[1]The continuous arrows indicate "belongs to family", e.g. Spanish belongs to the West-Iberian family, which belongs to the Ibero-Romance one. The dotted arrows give examples of languages in other sub-families (e.g. Russian belongs to the Slavic family, in a subfamily different from the West-Slavic).

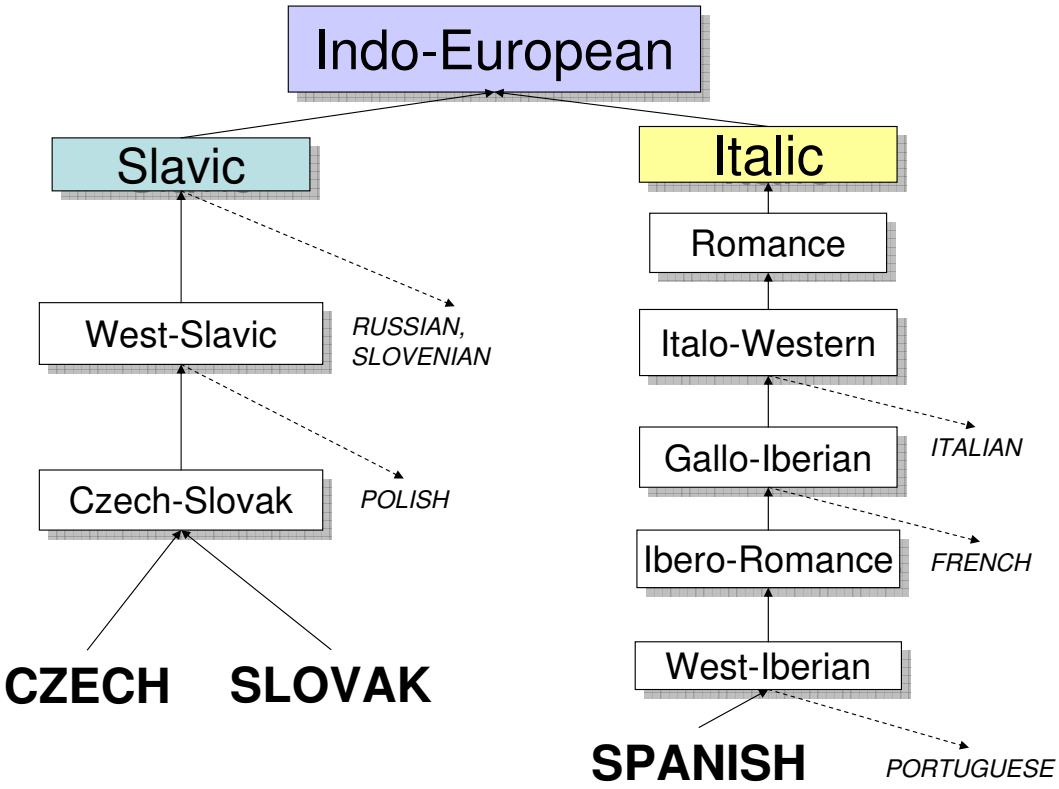


Figure 4.1. *Language family groups for Czech, Slovak and Spanish*

4.2 Related work

The term "cross-linguality" is used in many application domains in the field of computational linguistics, mainly to describe systems that can work employing several languages, and can be implemented at different levels. In the lexical level there are the natural language processing applications such as text retrieval, in which cross-lingual systems can search and rank documents written in a language different from the one in which the query is made (Fluhr et al., 1999; Martín-Valdivia et al., 2005). Cross-linguality is also applied to related areas such as question answering and text summarization (Radev et al., 2001; Ligozat et al., 2006). However, in these areas it is always necessary a translation between the involved languages, which must be carried out at some level of the text and query processing. Similarly, in

the semantic domain some intermediate representations are used to describe concepts and map them to the specific lexicalization of each language. For example, recent research is being done on how to map concepts to pictures to carry out multilingual web searches (Mihalcea and Leong, 2006).

When natural language is not written but spoken, cross-linguality can also be applied at the acoustic level. Traditionally, the most wide-spread systems that deal with cross-lingual acoustics are multilingual speech recognition systems (Schultz and Kirchhoff, 2006). These have been employed as components of speech-based interactive systems, as for example multilingual dialogue systems (López-Cózar and Araki, 2005), with which users can interact in different languages. Another example are speech-to-speech translation systems (Nakamura et al., 2004), which serve as real-time interpreters. Employing these systems, a user speaks to the telephone and the interlocutor receives the utterances translated to another language.

Acoustic cross-linguality has also been addressed from the point of view of resource sharing. The development of a speech recognizer is a very arduous and time demanding task. A large amount of data spoken by hundreds of subjects must be recorded and carefully annotated to get a representative set suitable for training an acoustic model. For example, the system used in our experiments for the speech recognition of Czech, which is described in Section 4.3, relies on phoneme HMMs that have been trained on approximately 50 hours of phonetically annotated speech provided by 700 speakers. For other languages the resources are often even much larger. Thus, collecting and annotating the necessary data generally requires many years of human effort. This is the reason why any possibility to share acoustic data between languages is very welcome in the scientific community. Similarly, a big effort is needed to create the linguistic part of a speech recognition system. For example, in the case of Czech it was necessary to collect a text corpus of around 4 GB, comprised of 456 million tokens with 1.9 different words and word-forms, to enable us to make a representative lexical analysis. It was shown that to obtain approximately 99% coverage of this language, at least 500,000 words had to be included in the general-purpose lexicon. This happens because of the inflective nature of Czech, which implies a big variety of word forms. However, the size of the needed vocabulary varies between languages. For example, English requires 20,000 words only to obtain the same order of coverage (Németh and Zainkó, 2003).

Frequently, these databases are gathered ad hoc by each developing team and are not freely available for the scientific community. This implies that in most cases, building a new speech recognizer requires obtaining all the necessary acoustic and linguistic resources right from the start. Thus, cross-lingual methods are an alternative to the complete creation of a new recognizer. Finally, sharing acoustics across languages has not only been used to create a speech recognizer from the resources of an already available one, but also to improve the performance of speech recognizers using models from another language. For example, some previous studies have shown that the recognition of Afrikaans can be improved by using additional English speech data (Nieuwoudt and Botha, 1999).

The usage of acoustic information across languages can be addressed following different strategies, which can be divided into two approaches: multilingual applications that can handle multiple languages simultaneously, and language adaptation where an existing recognizer is adapted to a new target language.

In the first approach, multilingual recognizers are capable of recognizing simultaneously several languages by sharing acoustic and/or language models. Multilingual acoustic models consist of either a collection of language-dependent acoustic models for each language, or a combination of language-independent acoustic models (Schultz and Kirchhoff, 2006). The main idea of the latter is to combine phonemes of several languages into one single acoustic model. To determine which phonemes of the different languages must be combined in the same category, some multilingual phoneme databases have been used, for example GlobalPhone (Schultz and Waibel, 1998). This technique relies on abstractions over the concept of *phoneme* in higher-order units such as meta-phonemes or archi-phonemes (Cahill and Tiberius, 2002), and assuming that phonemes in different languages can be grouped together similarly as allophones are considered inside the concept of phoneme.

The second approach carries out an adaptation of the "mono-language" speech recognizers to other languages. One possible method to do this is to create a mapping between phonemes. Alternative methods like word mapping (Bayeh et al., 2004), have been proposed. However, although using this method can sometimes lead to better results, it is more expensive and less practical than using phonemes. This happens because word mapping is less

prone to reusability, as it requires a complete translation of all possible words along with their different inflections. On the contrary, phonemes constitute a smaller set that can be used to automatically build any other higher-order construction such as words.

The basic idea of phoneme mapping is to establish a correspondence between phonetic units in the origin and target languages. Thus, the result depends on the phonetic similarity between both languages. This mapping can be done either automatically or by experts. The automatic procedure employs data-driven measures, which are frequently extracted from phoneme confusion matrices. The expert-driven approach is based on human knowledge about the languages being processed. Usually, International Phonetic Alphabet (IPA) [2] symbol tables are used for the different languages to determine the equivalent phonemes (Schultz and Kirchhoff, 2006), as IPA defines a unique representation of phonemes which can be used to compute equivalences between languages.

The mapping approach has been used because it was not in our scope to build a multilingual recognizer, but to use the Czech one with Slovak or Spanish utterances. Thus, the meta-phoneme concept was not effective for our purposes. To carry out the mapping, the expert-driven approach was chosen. The reasons for this are two. Firstly, although automatic mapping has the advantage of not needing human intervention and thus obtains more objective results, it requires considerable speech material for computing the similarities between languages. Even though this is not as much material as needed for building a full new recognizer for the target language, it makes the adaptation process be more costly. Secondly, results depend on a close match between the acoustics of both languages, on the used distance measure (Kumar et al., 2005), and on the recording conditions.

4.3 The MyVoice system and Czech speech recognizer

The Czech speech recognizer used in the experiments has been developed during more than a decade in Technical University of Liberec (Nouza et al., 2005). Its acoustic models are based on three-state left-to-right HMMs of

[2] IPA official web page: http://www.arts.gla.ac.uk/IPA/

context-independent speech units which are supplemented by several models of noise, with output distributions using at least 64 Gaussians. According to the application conditions, these models can be either speaker-independent (SI), gender-dependent (GD), or speaker-adapted (SA).

The recognizer's decoding module uses a lexicon of alphabetically ordered words, each represented by its text and phonetic form. This recognizer has been successfully employed for the development of the MyVoice and MyDictate systems (Cerva and Nouza, 2007). The former allows persons with non-functional hands to work with a PC in a hands-free manner by using several hundreds of voice commands. To do so, MyVoice interprets spoken commands into one or more basic actions; for example, pressing, holding and releasing key or combination of keys in a keyboard, moving the mouse cursor and clicking mouse buttons, starting executable programs, and printing sequences of characters. The latter is the first dictation program developed for the Czech language. It works with a very large vocabulary comprised of the 540,000 most frequent Czech words and it is primarily meant as a powerful aid for motor-handicapped users.

For the experimentation presented in the chapter, the MyVoice system has been employed and translated to Spanish and Slovak. MyVoice is structured in several command groups, each dealing with a specific task. For example, the group that controls the mouse is different from the one that deals with the keyboard, but they can be accessed easily from each other by voice commands. The size of these groups varies between 5 and 137 commands, where the largest group contains mainly the names of the alphabet letters, and of the keys on a PC keyboard, which makes recognition very difficult as the acoustic difference between them is very subtle. However, as a specific vocabulary is defined for each task, better recognition results are achieved when the commands are grouped. Besides, the grouping facilitates the interaction, as the user is aware of the valid words than can be uttered at any time. The MyVoice software is currently employed by 60 handicapped users in the Czech Republic, whose reports show that the word error rate (WER) typically lies between 1% and 2%, if the user does not have any speech disorder.

The development of MyVoice was motivated by the fact that there were no commercial tools of that type for Czech handicapped users. This is also the case for many other languages for which speech technology has not

been developed yet, and for which deploying such systems would require a big investment that would obtain little benefits due to the reduced target population. Therefore, we started to investigate the possibility to port software like MyVoice to other languages. Firstly focusing on the Slovak language, which is very similar to Czech. Secondly, making a more complex attempt to apply the same porting strategy to Spanish, which is a language that is acoustically and linguistically more different.

4.4 Cross-lingual adaptation

For the cross-language adaptation Slovak and Spanish texts were used along with an automatically generated Czech phonetic representation, which is discussed in Sections 4.4.1 and 4.4.2, respectively. The phonemes built for the Czech recognizer can be applied to recognize words in another language. To do this the Czech phonetic form was used, and the acoustic models of the words were constructed by concatenating the corresponding phoneme models. The translation from Slovak or Spanish text to the Czech phonetic representation was automatically done employing the mapping policies shown in Tables 4.1 and 4.2, respectively. For the phonetic representation a mapping of the IPA symbols to a set of ASCII symbols named Phonetic Alphabet for Czech (PAC) was used. We are aware of the fact that there are several encodings of the IPA alphabet such as XSAMPA. However, we have used PAC for two main reasons: firstly, because it has been established as a common base for speech processing research in the Czech Republic (Nouza et al., 1997). Secondly, because it has been successfully employed by the users of MyVoice, providing them with a straightforward language with which they can type their own phonetic transcriptions to customize the pronunciation of the voice commands.

As illustrated in Figure 4.2, the result of this process was a vocabulary with all the words that the speech recognizer accepted. For each word it contained a Slovak or Spanish text form and its Czech phonetic representation. This vocabulary was used by the Czech speech recognizer described in Section 4.3 as if it were a Czech vocabulary, thus not even a single line of code had to be changed in it. As a result, a user can utter a word in Spanish or Slovak, and the speech recognizer uses the Czech models to obtain the best Spanish or Slovak form candidate.

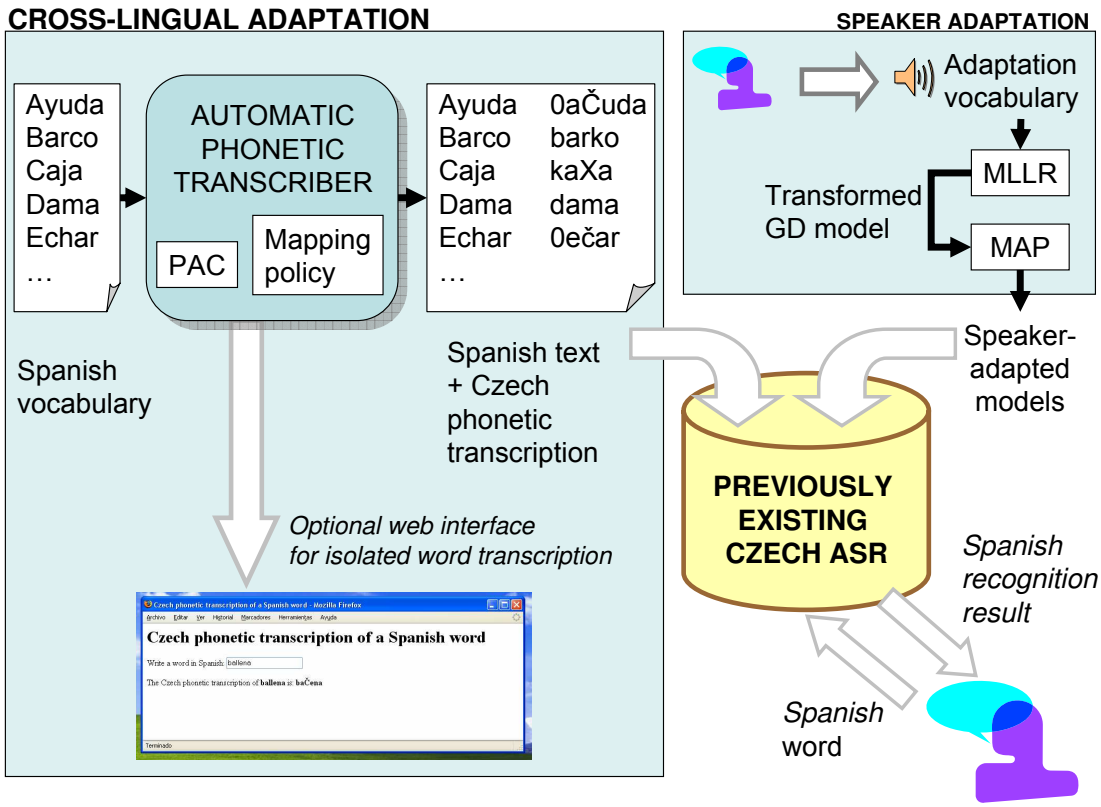


Figure 4.2. *Scheme of the cross-language adaptation procedure*

To optimize the performance of the recognizer, our proposal is to carry out speaker adaptation as a final step of the cross-language adaptation procedure. This way, the Czech models were tuned to better adapt to the pronunciation that each speaker had of every phoneme in the target languages. This step is not against the initial aim of cost-effective implementation, as it can be performed in a fast and straightforward manner by making the user read a short text when he first uses the recognizer (e.g. the first time he runs MyVoice). The proposal for speaker adaptation is described in Section 4.4.3.

4.4.1. Phoneme mapping between Czech and Slovak

As can be observed in Figure 4.1, Slovak and Czech belong to the same branch of Slavic languages. They share a large portion (about 40%) of their lexical

inventory and many of the remaining words differ slightly, either in spelling or pronunciation. In general, Slovak language sounds softer than Czech, which is caused by different phonotactics and a slightly different set of phonemes. The Slovak language uses four additional phonemes that do not occur in Czech: ''Í', 'l'', 'ŕ' and 'ŕ'. As shown in Table 4.1, these phonemes were mapped to the closest Czech phonemes. Concretely, the Slovak phonemes represented by characters 'Í' and 'l'' were mapped to the Czech phoneme 'l', Slovak 'ŕ' to Czech 'r', and Slovak diphthong 'ŕ' to the sequence 'uo'. The effect of this simplification on the performance of the adapted speech recognizer is very small, due to the similarity between the Slovak and Czech phonemes employed.

Czech phoneme	PAC	Slovak phoneme	Czech example	Slovak example
l	l	l, Í, l'	**l**ano	**l**ano, d**Í**hý, l'ud
r	r	ŕ	be**r**e	m**ŕ**tvy
uo	uo	ô	d**uo**	m**ô**že

Table 4.1. *Mapping of the Slovak phonemes that do not exist in Czech to the closest Czech ones*

4.4.2. Phoneme mapping between Czech and Spanish

The correspondence between Spanish and Czech phonemes was carried out by one Spanish native speaker and supervised by several Czech native speakers. As stated by Zgank et al. (2004), the accuracy of the correspondence depends on the number of phonemes present in each language and the similarity between them. It is difficult to find a consensus about the exact number of phonemes in the languages used in the experiments. This is especially the case for Spanish, as it has a high number of different varieties, even considering only the European branch and discarding the South American variants. However, in literature, Czech is generally considered to have around 40 phonemes, and Spanish around 20, which shows a big unbalancement between both languages.

The result of the mapping is presented in Table 4.2, which only shows the phonemes that were employed for the Spanish recognition. A complete list of Czech phonemes can be found in (Nouza et al., 1997).

Czech phoneme	PAC	Spanish phoneme	Czech example	Spanish example
a	a	a, á	ploch**a**	**a**nillo, **á**guila
b	b	b, v	**b**á**b**a	a**b**uela, **v**ino
č	č	ch	**č**ichá	**ch**arco
ch	X	g, j	**ch**udý	**j**aula, **g**ema
dž	Č	y, ll	rádža	llave, yema
d	d	d	je**d**en	**d**entro
e	e	e, é	l**e**v	**e**so, caf**é**
f	f	f	**f**auna	**f**auna
g	g	g, gu	**g**uma	**g**oma, **gu**isante
i, y	i	i, í	b**i**l, b**y**l	l**i**no, implí**c**ito
k	k	c, k, q	**k**upec	**k**ilo, **qu**eso, **c**asa
l	l	l	de**l**a	**l**ibro
m	m	m	**m**á**m**a	**m**adre
n	n	n	ví**n**o	vi**n**o
ň	ň	ñ	ko**n**ě	Espa**ñ**a
o	o	o, ó	k**o**l**o**	h**o**la, cami**ó**n
p	p	p	**p**u**p**en	**p**adre
r	r	r	be**r**e	a**r**co
s	s	s	**s**ud	**s**uelo
t	t	t	du**t**ý	**t**eja
u	u	u, ú, ü	d**u**še	l**u**na, **ú**til, ping**ü**ino
-	S / s	c, z	-	**z**umo, **c**ena
-	R / r	rr	-	**r**ueda, pe**rr**o

Table 4.2. *Mapping of the Spanish phonemes to the closest Czech ones*

As can be observed in the last two rows of the table, there are two Spanish phonemes that do not exist in Czech: $/\theta/$ and $/r/$, in IPA representation. For these phonemes two solutions have been studied. The first one was to use the nearest Czech phonemes: s and r, in PAC representation.

This mapping is not so unnatural, as the pronunciation of /θ/ as /s/ is even present in some varieties of Spanish, for example in Latin America and some areas in Southern Spain. The second option was to adapt these previously existing Czech phonemes to the Spanish pronunciation, by creating two new symbols, S and R, for the PAC table. Experimental results were very similar for both approaches, with a difference in accuracy of only 0.5%. The results reported in Section 4.6 were attained employing the second approach.

Additionally, as our system does not consider allophones, some sounds were ignored, affecting very slightly the experimental results. Moreover, there are differences in stress that make the recognition of Spanish more difficult. Czech words are always stressed in the first syllable, whereas the stress in Spanish varies between words. As noted by Carreiras et al. (1996), differences in stress in Spanish may be of importance not only for automatic speech recognition, but also for human listeners. However, the effect of stress is less important for isolated word recognition; even when some words that are differentiated in Spanish by their stress cannot be distinguished when translated to the Czech pronunciation, for example "este" and "esté".

4.4.3. Speaker adaptation

The approach proposed for speaker adaptation is a combination of the Maximum A Posteriori (MAP) (Gauvain and Lee, 1994) and the Maximum Likelihood Linear Regression (MLLR) (Gales and Woodland, 1996) methods for speaker adaptation, and is performed in two steps. In the first step, the mean vectors of the Czech gender-dependent models are transformed by the MLLR method. In the second step, these transformed values are used as priors for the MAP based adaptation. The main benefit of this approach is that the models that are not seen in the adaptation are well adapted by the MLLR method; while the MAP ensures that the parameters of the models with a lot of adaptation data can converge to the values of the theoretically best speaker-adapted model.

To carry out the speaker adaptation a 614 words vocabulary was used. It was comprised of a list of the most frequent words in each language (covering all phonemes), along with MyVoice commands. This decision was taken from experimental results, which showed that in most cases, misrecognitions occurred for monosyllabic and short words. Besides, these words, which for

example in Spanish generally are pronouns, determinants and prepositions, are usually the most frequent. Thus, misrecognition of these words had a big impact in the computation of accuracy. Concretely, a 44.9% relative improvement was achieved for Spanish when using this vocabulary for adaptation instead of the 432 MyVoice commands only. Additional experiments showed that this improvement was better than the one obtained using the same amount of phonemes for adaptation, but extracting them from words selected from newspapers instead of using the most frequent words of each language; even when both word lists considered all the phonemes of each language, and these were covered in identical proportions.

4.5 Experimental set-up

Several experiments were carried out with the main objective of testing the viability and performance of the proposed cross-lingual adaptation approach. Additionally, it was measured the impact of several factors in the performance of the employed method, such as the usage of different user adaptation strategies, the size of the recognition dictionary, and the number of words considered for testing, as shown in Figure 4.3.

Firstly, the MyVoice commands were translated into Slovak and Spanish, to measure the performance of the cross-lingual adaptation for a command-and-control application. The speakers used MyVoice to control a PC by spoken commands while they carried out their daily activities. Thus, they were not provided with a specific list of commands to utter. This way, the results were attained from flexible and natural interaction with the system.

As it was described in Section 4.3, the valid vocabulary of MyVoice is restricted at each step to the list of commands in the current group (vocabulary ranging between 5 and 137 commands). As commented before, the largest group was designed for the recognition of spoken characters. This is a very challenging task, mainly due to the acoustic similarity of the characters, which makes them highly confusable. Thus, although the grouping of commands considerably reduces the vocabulary that can be recognized at each time, speech recognition during natural interaction with MyVoice is not trivial, as it can involve the recognition of spoken characters.

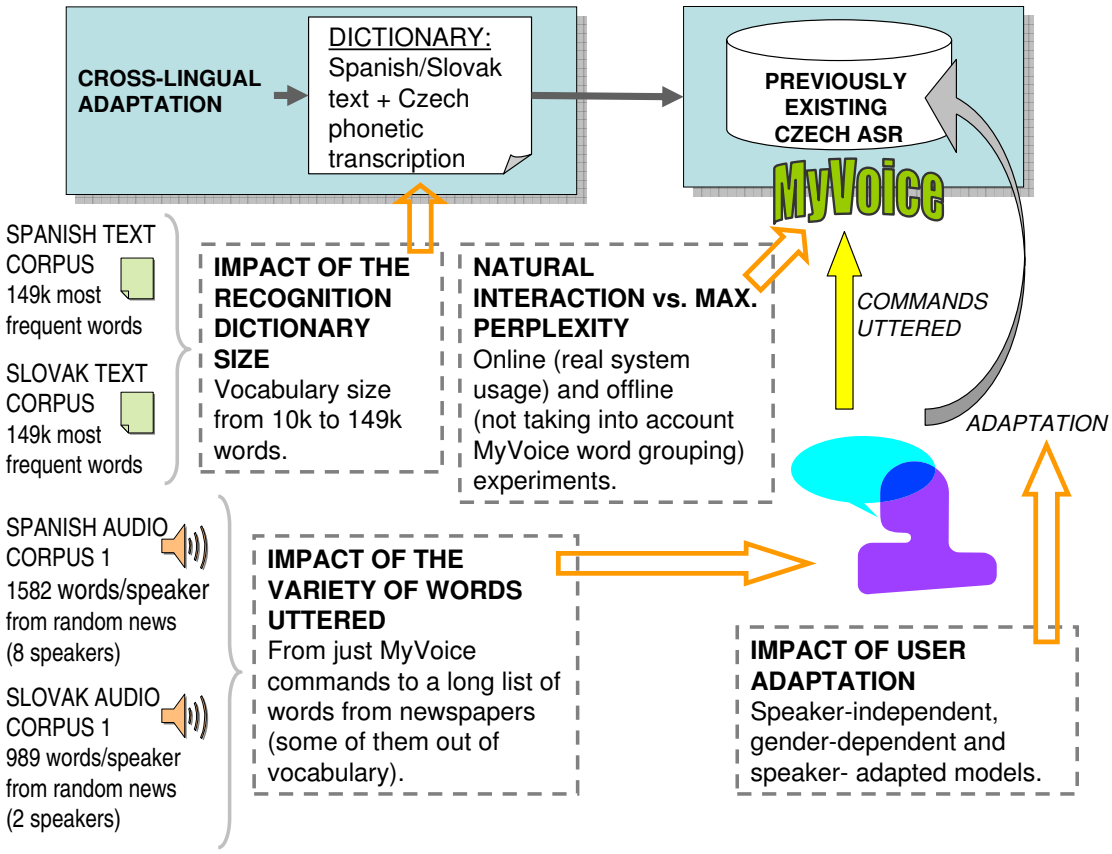


Figure 4.3. *Outline of the experimental set-up*

To obtain meaningful results from the different speaker models regardless of the groups visited during the interaction, additional experiments were carried out employing the whole MyVoice vocabulary (432 commands). In these experiments the task perplexity was always 432, given that after each command, any word could be uttered.

We also wanted to corroborate that the results obtained from the interaction with MyVoice could be attainable in situations where the accepted vocabulary was larger. Thus, MyVoice commands were extended with a list of the 149k most frequent Spanish and Slovak words, respectively. The two dictionaries were collected from Spanish and Slovak newspapers and contained all the word forms, not only the lexemes. The words were automatically annotated with their appearance frequency and sorted from most frequent to least. In these experiments recognition dictionaries of sizes ranging from 10k

to 149k words were used, which were subsets of the described dictionary, from the most frequent word up to the desired number of words. Our aim was not to build a Slovak or Spanish dictation system, but simply to check whether the cross-lingual approach would also work efficiently for a more complex task with increasingly larger vocabularies. For this purpose, a vocabulary of 149k words is larger than those usually employed in previous cross-lingual studies. For example, Zgank et al. (2004) only varied their vocabulary size from two to several thousand words, whereas Nieuwoudt and Botha (1999) employed 60 sentences as the maximum for their test purposes.

At the same time that the list of MyVoice available commands was extended, it was also augmented the vocabulary employed to test the system. To do this, news from Spanish and Slovak newspapers were randomly selected. They were different from the ones used to collect the recognition dictionaries, and belonged to different categories, namely: politics, economy, culture and sport. Eight Spanish native speakers (four male and four female) aged 23 to 60, and two Slovak native speakers (one male and one female) aged 22 and 24, recorded the isolated words. The experiments with the two Slovak speakers using MyVoice to carry out their daily activities, showed that the WER was almost as small as for native Czech speakers (only 2.5%), and that the space for any further improvement was very small. Therefore, a decision was taken to not to carry out any additional experiments with other Slovak speakers, highlighting that the main contribution is the adaptation to Spanish. Concretely, 1,582 words were recorded by each Spanish speaker, and 989 by each Slovak speaker. For discussing the experimental results we used the average performance values over the number of words for each language, and for all speakers. To obtain results that can be significant for the usage of the proposed method in real conditions, the corpora were not recorded in a closed laboratory environment, but in each speaker's PC in order to reflect the real noise conditions in which the MyVoice system would be used.

Furthermore, in order to study to which extent speaker adaptation allowed us to attain better recognition results, experiments were carried out with speaker-independent, gender-dependent, as well as speaker-adapted models. The improvements achieved by each adaptation step were compared and studied taking into account the impact of out-of-vocabulary (OOV) words and the size of the recognition vocabulary.

4.6 Experimental results

4.6.1. Interaction with MyVoice

In the first experiments the users employed MyVoice to control their PCs in order to carry out their daily activities. The experiments were performed both online and offline. The online results were extracted from natural interactions of the users with the MyVoice system. Thus, as commented before, the perplexity of the recognition task varied from 5 to 137, as the valid vocabulary was comprised of the words in the current command group. In the offline case, the same utterances recorded during the online experiments were used. However, recognition was carried out offline using as valid vocabulary all MyVoice commands. Hence, as all the commands could be uttered any time, the recognition perplexity was 432 (the number of MyVoice commands).

The experimental results are shown in Table 4.3. It can be observed that WER was lower for the online experiments because the vocabulary size was smaller. When using speaker-adapted models, the relative improvements achieved were 24.1% for Slovak and 28.3% for Spanish for the online experiments; whereas they were 46.65% and 56% respectively for the offline experiments. This shows that speaker adaptation considerably reduced the difference in performance observed for the online and offline recognition, as it caused a remarkable improvement in the offline recognition results; which were comparable to the ones obtained in the online experiments after speaker adaptation (around 2% WER for Slovak and 4% for Spanish).

Language	Experiment	Gender-dependent	Speaker-adapted
Slovak	Online	2.9	2.2
	Offline	4.6	2.5
Spanish	Online	6.0	4.3
	Offline	10.0	4.4

Table 4.3. *WER [in %] for the command-and-control task*

The experiments with Slovak showed a WER almost as small as for native Czech speakers (only 2.5%). The results with Spanish were much better than initially expected. In fact, they differed in less than 2% compared with the ones that could be achieved recognizing the Czech language.

4.6.2. Impact of speaker adaptation

The good results obtained using speaker adaptation during the natural interaction with MyVoice, encouraged us to study to which extent the speaker-adapted models convey an improvement in the proposed cross-lingual approach. To test the performance of the adapted recognizers the Spanish corpus (12,686 words extracted from newspapers) and the Slovak corpus (1,978 words) were used, which were described in Section 4.5. Additionally, we used a 10k words vocabulary for recognition instead of the small vocabulary comprised of 432 commands. As can be observed in Figure 4.4, speaker-independent models yielded a WER of 47% for Slovak and 55.8% for Spanish. These results were improved by using gender-adapted models only in a 7.23% relative for Slovak and 2.15% for Spanish. However, speaker adaptation yielded a 17.6% relative improvement with respect to the speaker-independent models for Slovak, and a remarkable 40.8% relative improvement for Spanish. Most of the recognition errors for Slovak were due to OOV words. Thus, as the objective of this experiment was to measure the impact of speaker adaptation in the proposed cross-language approach, regardless of the dictionary and utterances used for recognition, the recognition results were computed without considering the OOV words. As shown in Figure 4.4, WER decreased for both languages when OOV words were not considered. Concretely, for Slovak the decrement was of 20.9% absolute for the best case. Regarding speaker adaptation, a 54% relative improvement with respect to speaker-independent models was achieved for Spanish, and a 38.2% in the case of Slovak, obtaining for both languages WERs around 20%.

As Slovak is a language very similar to Czech, initially the proposed method attains accuracies around 70% (29% WER) for this language. Hence, speaker adaptation for Slovak only improves accuracy by an absolute 11.1% (38.2% relative) with respect to using speaker-independent models. However, for a language with a very different origin such as Spanish, speaker adaptation enhances the adapted recognizer substantially. The experiments showed that 26.4% absolute improvement (54% relative) can be achieved, with accuracy rates that are only 4.6% worse than the ones obtained for the Slovak language. The proposed cross-lingual approach in combination with speaker adaptation yielded accuracy rates around 80% (17.9% and 22.5% WER) for Slovak and Spanish.

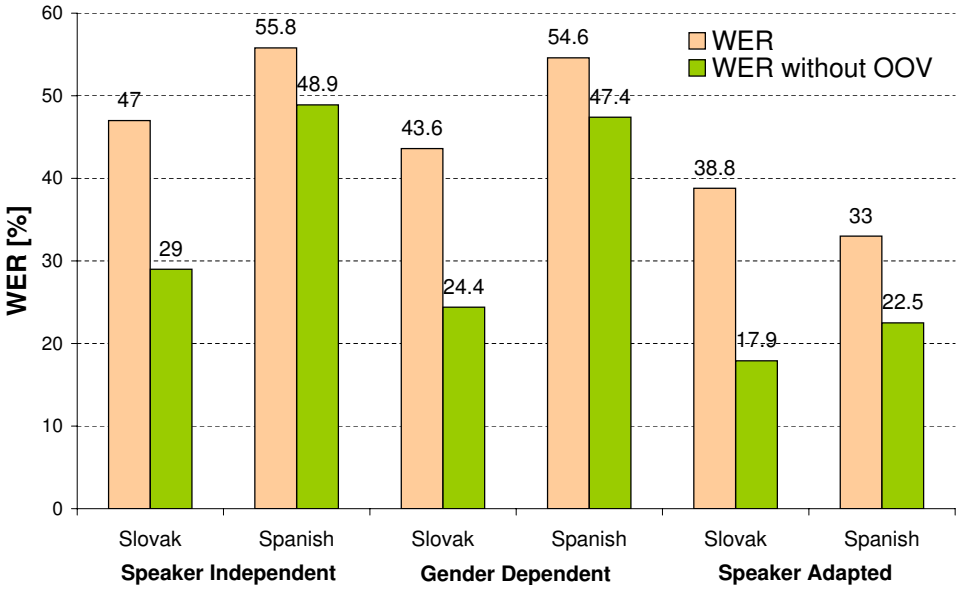


Figure 4.4. *Effect of the adaptation technique on the performance of the adapted recognizers*

4.6.3. Effect of the size of the recognition dictionary

Finally, we were interested in studying to what extent the experimental results could be affected by increasing the size of the recognition vocabulary up to 149k words. This size is inversely proportional to the number of OOV words that can appear during the recognition process, and at the same time it is directly related to the probability of an uttered word being acoustically similar to others in the vocabulary.

The number of OOV words decrements drastically when the recognition vocabulary is very large. Hence, WER tends to decrease when such a dictionary is employed. As can be observed in table 4.4, WER decreases 14.7% relative for Slovak when employing a vocabulary comprised of 149k words, compared to 10k, and 15.4% for Spanish.

	10k	46k	85k	149k
Slovak	38.8	27.3	26.0	24.1
Spanish	33.0	28.4	28.0	27.9

Table 4.4. *Effect of dictionary size on WER [in %] taking into account OOV words and speaker adaptation*

The experimental results also showed that the smallest WERs were obtained after speaker adaptation, as observed in figure 4.5. The objective was to check whether other languages could be efficiently recognized using the Czech recognizer, without using a specific vocabulary for recognition. Hence, Figure 4.5 only shows the results obtained by ignoring the OOV words.

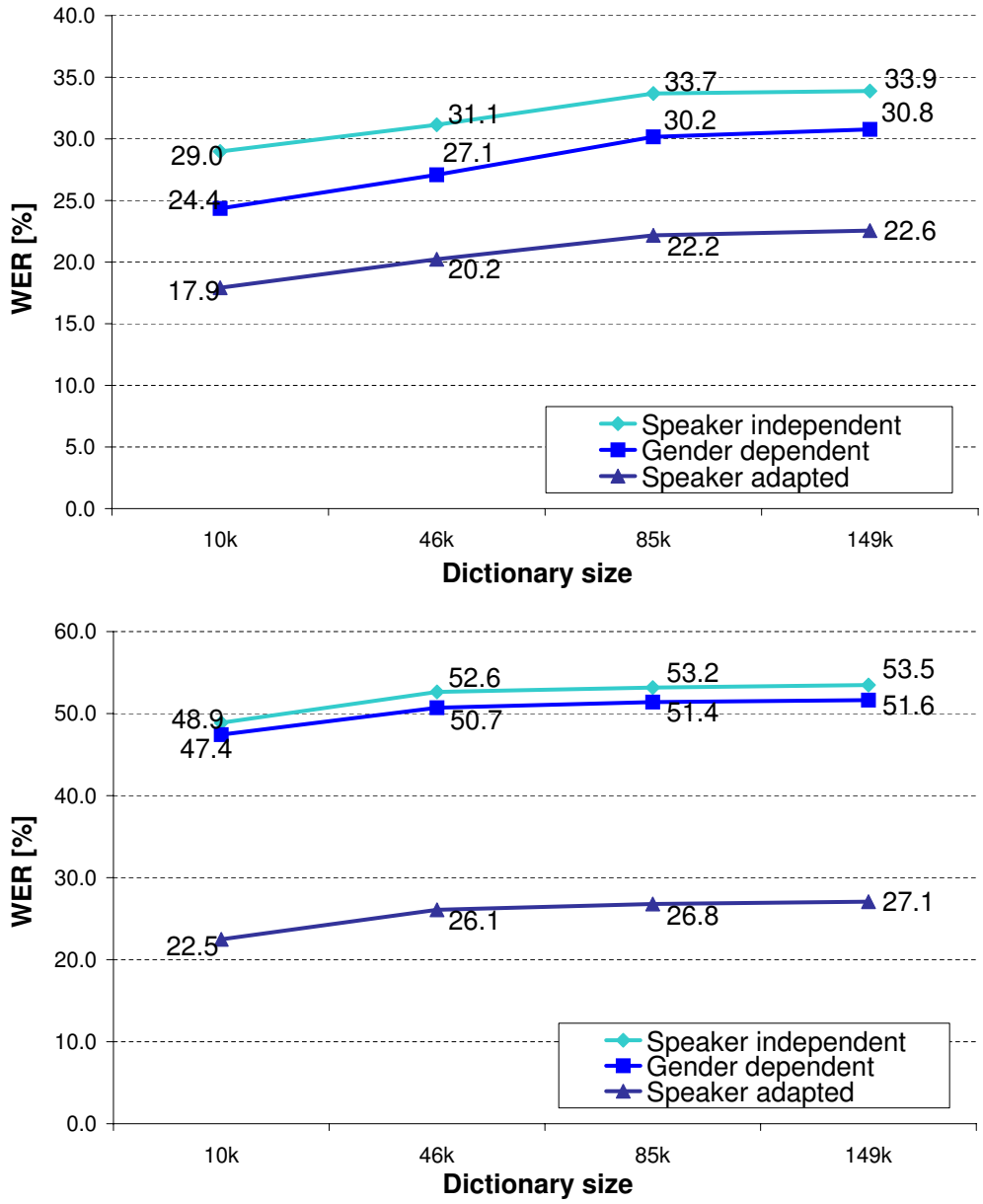


Figure 4.5. *Performance of the Slovak (above) and Spanish (below) adapted recognizers*

As can be observed, each adaptation type conveys an improvement. This improvement is larger for Spanish, for which it is necessary to use speaker-adapted models to similar results to those obtained for Slovak. As shown in table 4.5, the relative reduction of WER from speaker-independent to speaker-adapted models, decreases when the accepted vocabulary is larger. This is due to an increment of the probability to find acoustically similar words in the dictionary. However, as can be observed in Figure 4.5, for increasingly larger recognition dictionaries it was found that the WER tends to establish around 23% for Slovak and 27% for Spanish. This shows that the proposed cross-lingual approach yields accuracy rates which are around 70% when adapting Czech to Spanish, i.e. two languages with very different origin. Whereas the accuracy rates are around 80% when adapting Czech to Slovak, i.e. two very phonetically similar languages.

	10k	46k	85k	149k
Slovak	38.2	35	34.2	33.4
Spanish	54	50.4	49.6	49.4

Table 4.5. *Relative WER reduction [in %] yielded by the adaptation to speakers*

4.7 Conclusions

It has been presented in this chapter a cross-lingual adaptation of a previously-created Czech speech recognizer to Spanish and Slovak. Phonetic cross-linguality is a research area that is gaining increasing interest, especially because it enables resource sharing between languages and thus represents a feasible way of developing systems for minority languages or dialects. As rapid and low-cost development of speech-based systems is essential to foster portability, cross-language adaptation has arisen as one of the main challenges in the area (Gao et al., 2005). However, the state-of-the-art systems are based on complicated and very effort and time demanding linguistic and phonetic studies. Small effort has been devoted so far to study methods to carry out the cross-linguistic adaptation of a speech recognition environment (i.e. acoustic, lexical and linguistic models) in a cost-effective way.

It has been demonstrated that the adaptation of a speech recognizer to another language can be carried out in a straightforward way, employing a mapping between phonemes, and enhancing it with language and speaker adaptation procedures. Moreover, it has been shown that the proposed adaptation method can be used not only with phonetically similar languages, such as Czech and Slovak, but also with languages from very different families, like Czech (Slavic, Czech-Slovak) and Spanish (Italic, West-Iberian).

Several experiments have been carried out using MyVoice, a speech-based application designed for Czech handicapped people. Cross-lingual porting of voice operating systems for such a small group of target users, requires an investment that hardly can be paid back. However, our experimental results show that for a task involving a vocabulary of 432 commands, a 95.6% performance (4.4% WER) can be attained for Spanish and 97.5 (2.5% WER) for Slovak, employing the proposed cost-efficient procedure. Besides, for vocabularies up to 149k words, the proposed scheme yields around 72.9% accuracy (27.1% WER) for Spanish and 77.4% (22.6% WER) for Slovak.

Ce n'est pas assez de compter les experiences, il les faut poiser et assortir: et les faut avoir digerees et alambiquees, pour en tirer les raisons et conclusions qu'elles portent.

Michel de Montaigne, Essais

5 Field evaluation of spoken dialogue systems

5.1 Introduction

As stated in the introduction of the Thesis, dialogue systems are becoming increasingly attractive for a wide range of applications (McTear, 2004; López-Cózar and Araki, 2005; Wahlster, 2006). In order to minimize costs and optimize results, there is a need for standard methods, architectures and criteria to test, compare and predict the performance and usability of the systems. Several initiatives have arisen since the late 80s to establish these methods. In the USA, the main funding institution for this kind of research is DARPA (Defense Advanced Research Projects Agency), with their project COMMUNICATOR (Walker et al., 2002a), which was aimed at cost-effective development of multimodal dialogue systems. This was achieved by using different plug-and-play components which were evaluated paying special attention to user satisfaction maximization. In Europe, the major institutions concerned with evaluation of dialogue systems have been COCOSDA (Coordinating Committee on Speech Databases and Speech I/O Systems Assessment), which focuses on obtaining corpora that can be shared to study evaluation criteria[1], EAGLES (1996) and DISC (1999). These last two international projects established some best practice guidelines for the development and evaluation of dialogue systems, both at system and component level.

[1] http://www.cocosda.org/

These research efforts have successfully established a common background of criteria for quantitative evaluation. However, there is still no systematic understanding, nor consensus on the criteria that must be taken into account to optimize the usability of dialogue systems. Some projects have tried to address the problem of predicting system usability and user satisfaction from measurable performance criteria. This is the case of the PARADISE framework (Walker et al., 2000a), which has become one of the reference frameworks for system evaluation.

Because of the complexity and effort demanded by the application of the PARADISE framework, many approaches in the literature apply qualitative and quantitative measures separately. For example, Hartikainen et al. (2004) propose a methodology for subjective evaluation that has been used for evaluating the MUMS Multimodal Route Navigation System (Hurtig, 2004). Recently, the VIrtual CO-driver system (Geutner et al., 2002), the MASK multimedia service kiosk (Lamel et al., 2002) and the SAMMIE dialogue system (Becker et al., 2006), have been also evaluated only subjectively. Other authors, for example Robinson et al. (2006), evaluate their systems both with instrumentally-derived measures and quality judgments, but without establishing links between the different evaluation measures employed. In this chapter empirical results are obtained on the relationship between both types of criteria from the evaluation of our spoken dialogue system. This is done via correlation studies, which we believe are a reliable method that can be applied to both whole system and component level evaluation. However, when the statistical studies are carried out over a large number of metrics, there is a possibility that some of the findings are due to chance, and thus reliability and significance studies are also reported. This method has been applied successfully for the evaluation of other dialogue systems, e.g. BoRIS (Möller, 2005), yielding some interesting relationships between evaluation criteria.

However, results in the literature are usually based on restricted laboratory interactions, in which some users are asked to interact with the system in accordance with predefined scenarios. In some cases the users are also given evaluation questionnaires in which they express their personal opinion about different interaction aspects. The main disadvantage of this method is that the scenarios may differ from the tasks that a user would have selected in a non-predefined interaction. In contrast, field evaluation requires real

users interacting with the final system in their appropriate environments. Although as stated by Bernsen and Dybkjaer (2000), field tests can fail to be representative of the full functionality of the systems, we believe they offer the most realistic results and cover real user motivations. Field evaluations are not repeatable as the interaction context is highly variable. This is also their main advantage as they gather results from different users (difference in gender, voice, knowledge, experience using the system), who talk on different devices (mobile phones, usual phones or PCs), and in different environments (different noise conditions). As the results obtained from field tests are robust to this heterogeneity, they are more relevant at predicting the real behaviour of the systems. The contribution of the Thesis to the state-of-the-art system evaluation relies on obtaining new empirical evidence by means of a field study carried out employing our spoken dialogue system.

The chapter is structured as follows. Section 5.2 presents an overview of the main evaluation trends that can be found in the literature. Section 5.3 describes the computation of the evaluation criteria, distinguishing between interaction parameters and quality judgments. Section 5.4 presents the statistical studies carried out, whereas Section 5.5 discusses the experimental results obtained. Finally, Section 5.6 presents the conclusions obtained.

5.2 Related work

Evaluation of dialogue systems has been used in the literature for a wide range of purposes, for example, measuring the system's performance, comparing a system with its previous versions to measure the adequacy of changes, comparing different systems and predicting system behaviour.

Regardless of its purpose, evaluation can be carried out using "glass-box" or "black-box" approaches. The former permits access to internal details of the system to measure their contribution to the overall performance. The latter, treats the system as a black box so that the evaluation is based uniquely on the response of the system to the different user inputs. In both cases evaluation can be developed over a complete system or over an individual component. Usually, evaluation is carried out at the component level, the main working areas being the assessment of the speech recognizer, the speech and natural language understanding components, the dialogue manager and the speech output components.

Speech recognition performance is usually assessed in terms of automatically generated measures that calculate the number and importance of the recognition errors. Most of these measures are generalized in their use, like for example word error rate (WER) and word accuracy (WA), which are complementary in their use. WER is defined as the number of incorrectly determined words (calculated as the sum of the number of substitutions, deletions and alterations made by the recognizer) divided by the total number of words, and WA can be calculated as $1 - WER$. Apart from the correct functioning of the speech recognition engine, when employing glass-box approaches, internal components of the speech recognizer can also be studied like for example language models, lexicons or phonetic models. Additionally, having information about some of the recognizer components, can serve for prediction studies of the final recognition performance. For example in (Persia et al., 2007), they use performance of the source separation algorithm as a prediction of the speech recognition success in noisy environments. A best practices summary for evaluation of speech recognizers can be found in (Lamel et al., 2000a).

Despite the generalized use of the measures, the assessment of different speech recognition systems differs in several experimental conditions which make comparisons very difficult. Some of these influencing factors are the vocabulary characteristics (e.g. whether it is isolated or continuous, the vocabulary size, or the phonetic similarity of the words), the acoustic environment (e.g. noise levels), the transmission characteristics (e.g. transmission errors or signal levels) and the speaker characteristics (e.g. age, gender or cultural background). All these factors have to be taken into account when trying to compare the evaluation results of different recognizers, which makes this task very complicated. Recently, some studies have focused on how to overcome these difficulties by creating corpora that can be shared between the scientific community. The objective is to provide a common testbed for evaluating and subsequently comparing speech recognition and enhancement algorithms, like for example the NOIZEUS speech corpus (Hu and Loizou, 2007).

For the assessment of speech and natural language understanding engines, some measures similar to WER have been used (Gupta et al., 2006). For example slot error rate (defined as the number of incorrect slots divided by the number of slots), the update precision (defined as the number of

correctly updated slots divided by the number of updated slots) and the concept error rate (CER), which is defined as the number of incorrect slots divided by the number of filled slots. A more detailed list of metrics can be found in (Higashinaka et al., 2004). However, evaluating understanding modules requires a higher involvement of experts in judging whether the results are correct, and thus there is a need for a higher level of quality judgment measures. One of the seminal studies in this area was developed inside the TSNLP (Test Suites for Natural Language Processing) project (Lehmann et al., 1996). Furthermore, evaluating natural language understanding components is very domain-dependent, as they strongly rely on the semantics of the task and the detail to which it is done (i.e. the number of concepts which are used), and thus it is difficult to compare evaluation results between systems.

Comparison is even more complicated in the case of the speech output components, because their evaluation is almost entirely done using judgment measures about their quality, thus obtaining highly subjective evaluation results. Measures used mainly describe the quality and/or naturalness of the speech sounds, as well as the comprehension of the system message by the user. The results vary greatly depending on the users, that is why most studies are centred in specific population sectors, like for example the elderly in (Lines and Hone, 2002). A comprehensive list of intelligibility, prosody and overall quality tests to evaluate speech synthesized outputs can be found in (Gibbon et al., 1997).

Dialogue management is usually evaluated in terms of the quality of the user-system interaction: adequacy of system responses, feedback strategies, duration of the dialogue, average number of turns, adequacy of the initiative, adequacy of confirmation strategies or ability to solve misunderstanding situations. The DISC (1999) work group proposes 6 main groups that gather the criteria to be taken into account when evaluating dialogue managers: correct management of knowledge about the current dialogue context, mapping from the semantically significant units in the user's most recent input, analysing the user's specific contribution, generating of output to the user, specific issues of dialogue management evaluation (e.g. feedback strategies) and global issues of dialogue system evaluation (e.g. time for task completion). It is difficult to separate the evaluation of the dialogue manager from the rest of the system, specially in quality judgments, this is

why some authors propose to use manually corrected input to the dialogue manager to be able to have some "gold standard" to represent the dialogue manager behaviour in isolation. This baseline can be used subsequently for comparison with real situations in which the dialogue manager is affected by errors in the recognition or understanding modules (Roque et al., 2006a). Additionally, the dialogue manager makes intensive use of domain knowledge and has to be able to build and adapt to the interaction context. Thus, some authors like (Hanna et al., 2007), have studied how to carry out evaluation of dialogue managers in terms of modification of the domain-specific expertise and maintenance and reuse of the already existing context knowledge and discourse management behaviour.

There is not a consensus in the literature about the terminology to use for categorizing all the described evaluation criteria. Traditionally, authors have differentiated between objective and subjective evaluation criteria. The former takes into account measures computed from system performance features such as word error rate (WER). The latter considers measures that judge some property, for example intelligibility of the synthesized speech. This notation has been widely used in previous studies, for example, Larsen (2003), Minker et al. (2004b) and Robinson et al. (2006). However, as argued by Möller (2005), human subjects are always involved in determining the systems' performance. In the so-called objective measures human expert evaluators are often used, for example to calculate WER, experts have to compare real user input with the recognizer output). Thus, Möller (2005) proposes to differentiate between quality judgments (subjective), interaction parameters (which can be instrumentally measured or expert derived) and quality predictions (which can be instrumentally derived). This chapter will be focused on the first two categories.

There have been several attempts to create a full list of criteria to be used for evaluation by employing interaction parameters, quality predictions and quality judgments. For example, Dybkjaer and Bernsen (2000) propose a list of 15 criteria to guarantee system usability: adequate use of modalities, accurate input recognition, flexibility of the accepted vocabulary, system voice quality, adequate response generation, adequate domain coverage, and user satisfaction, among others. The Expert Advisory Group on Language Engineering Standards (EAGLES, 1996), proposed quantitative (e.g. system response time) and qualitative measures (e.g. user satisfaction),

that were applied and interpreted following an innovative framework. This framework provided guidelines on how to carry out the evaluation and how to make results available in such a way that they could be easily interpretable and comparable. In the DISC (1999) project there were other best practice guidelines that completed the EAGLES proposal using life cycle development methodologies. Other authors have focused on how to obtain and study speech corpora to compute evaluation measures. These are frequently large corpora extracted from system usage, or from human-to-human dialogues. In the latter case, human behaviour can be used as a baseline to compare with the system behaviour (Paek, 2001). For example, the project EVALDA (Devillers et al., 2004) focuses on evaluation campaigns that consider various aspects of natural language interaction. One of them is the MEDIA campaign, which evaluates the interaction between users and dialogue systems. Their evaluation methodology employs test sets obtained from real corpora along with the commonly used evaluation criteria. Degerstedt and Jönsson (2006) proposed the LINTEST tool to carry out evaluation of dialogue systems using the JUNIT corpus. A very detailed review of the most relevant efforts on generalization of evaluation criteria and practices can be found in Dybkjaer et al. (2004) and in López-Cózar and Araki (2005), whereas Möller et al. (2007) present a review of the de-facto criteria extracted from all these studies and an example of their usage to evaluate a particular dialogue system.

As commented above, PARADISE (Walker et al., 1998b) is the most widely embraced evaluation method proposed so far to specify the relative contribution of various factors to the overall system performance. This method models performance as a weighted function of: task success (exact scenario completion), dialogue efficiency (task duration, system turns, user turns, total turns), dialogue quality (word accuracy, response latency) and user satisfaction (sum of TTS performance, ease of task, user expertise, expected behaviour, future use). More recently, PARADISE has been used to develop models of user satisfaction prediction, again based on the weighted linear combination of different measures (Walker et al., 2000b). The goal of this evaluation method is to maximize user satisfaction by maximizing task success and minimizing interaction costs as shown in Figure 5.1. These costs are quantified using different efficiency and quality measures. The weights of each measure are computed via a multivariable linear regression consid-

ering user satisfaction as the dependent variable and task success, efficiency and quality measures as independent variables. Recently, the PARADISE framework has been enhanced to enable evaluation of multimodal dialogue systems. For example, it was used in inside the SmartKom Project, creating the so-called PROMISE framework (Beringer et al., 2002).

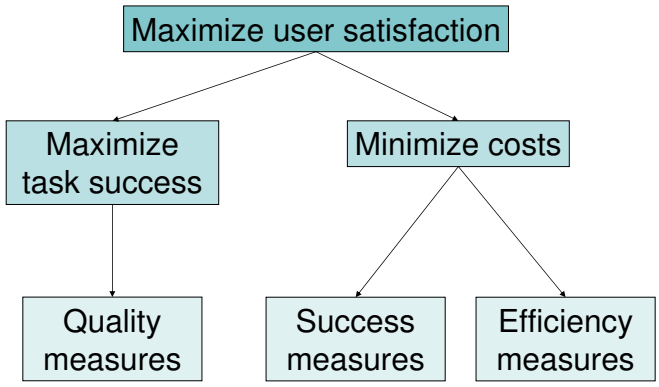


Figure 5.1. *PARADISE architecture model*

The application of PARADISE to evaluate a dialogue system requires dialogue corpora extracted from controlled experiments in which subjects have to evaluate satisfaction on a scale after they have interacted with the system. This approach has been successfully used for evaluating and comparing eight COMMUNICATOR systems (Walker et al., 2002b,a), firstly in controlled laboratory experiments, and secondly in a less restricted context where the systems were accessible on the phone. Strictly, this second evaluation was not an open field study because the authors had control over the users, who were specifically recruited and assigned to the different systems. Nevertheless, the tasks they had to complete were not predefined in all cases. A similar approach was employed in the ARISE project (den Os et al., 1999), where evaluation was based on the responses of subjects who either called a dialogue system from home or interacted with it in the laboratory. In either case, the tasks to be carried out by users where predefined (Sanderman et al., 1998).

It can be also found in the literature a distinction between "internal" and "external" tests regarding whether they were carried out by users from the development team of the dialogue system (internal evaluation) or by users who did not have any previous knowledge about the system (external evaluation). However this is not equivalent to "field" vs. "laboratory"

studies distinction, as external tests may involve using predefined scenarios. For example, Rajman et al. (2004) propose a Rapid Dialogue Prototyping Methodology to produce, for any given application, a quickly deployable dialogue-driven interface which can be later enhanced through an iterative Wizard-of-Oz process. To refine the dialogue models developed using this methodology, the authors propose to use an internal and an external test. The internal test is used to further adapt the prototype and its successive modifications. The external test is employed for the final evaluation of the resulting dialogue interface. In both cases the evaluation is carried out in the form of a satisfaction questionnaire which is submitted to the users after they have interacted with the prototype, on the basis of a set of predefined scenarios involving specific contexts for a restaurant search.

To study the implications of using field tests, some authors have focused on non-restricted evaluation studies. This is the case of the Let's Go system (Raux et al., 2003), which was evaluated using interactions of real users that phoned the system to get information about bus schedules. The evaluation was carried out by reporting results of interaction parameters (Raux et al., 2006). Unfortunately, although these parameters are relatively easy to compute, they do not provide sufficient information on quality. Qualitative judgments, on the other hand, are difficult to extract and compare when they are related to subjective opinions. Only in a few cases, performance parameters which can be measured quantitatively are also able to express quality. Our work focuses on using both quantitative and qualitative de-facto standard measures (Möller et al., 2007) in a field study, to evaluate our spoken dialogue system, which is described in Chapter 2. Our main objective is to empirically obtain relationships between these measures by employing statistical significance studies. Similar methods have been widely used in the area of systems' acceptance, more specifically for predicting the adoption of new technologies, e.g. in risk studies by investing companies. One of the most used models is the Technology Acceptance Model, which relates several user judgments criteria with the final adoption of the technologies by users (Legris et al., 2003). However, no quantitative parameters are considered in this model. In the area of dialogue systems, very few authors have exploited correlation studies to measure such relationships, for example Litman and Pan (2002), Möller (2005) and Schiel (2006), who applied it to controlled laboratory studies.

5.3 Evaluation criteria

The UAH system evaluation was carried out both with interaction parameters and quality judgments. Interaction parameters were employed to measure the system performance (e.g. number or errors made by the speech recognizer), and the dialogue course (e.g. duration of the dialogue or number of turns). These measures allowed to carry out different studies about performance and reliability of the system as well as discovering interaction points which can be improved. Although interaction parameters are a good indicator of the quality of the evaluated interaction, they do not necessarily provide reliable information about user satisfaction (López-Cózar and Araki, 2005). Thus, it is necessary to carry out a qualitative judgment evaluation to register users' opinions about these aspects of the interaction. In the experiments presented in the chapter, the subjective evaluation was carried out by employing user tests.

5.3.1. Interaction parameters

To compute the values for the interaction parameters, the UAH corpus has been used. This corpus consists of 85 dialogues and 422 user turns, with an average of 5 user turns per dialogue. Each dialogue was automatically annotated with two timestamps, corresponding to the call starting and ending times respectively. Each user utterance was stored in WAV format along with information about the recording starting time, the previous system turn, and the speech recognition result, which included confidence scores attached to the recognized words.

Then, it was manually annotated whether each utterance was correctly understood by the system, regardless of the speech recognition errors. For example, if in response to the system prompt: "What type of information do you want?", the user answered: "I want information about a subject", but the recognition result was: "Information about subjects", there are three deletions and one substitution. Regardless of these errors, the utterance was correctly understood by the system, as the semantic values returned by the speech recognition grammar were correct. Hence, the annotator tagged the utterance as "correctly understood".

At the dialogue level, the annotator registered the gender of the speaker, whether the dialogue was complete (i.e. whether the user did not hang up before finishing the dialogue) and whether the dialogue was successful. As it was a field study, there were no predefined tasks for the users to accomplish. Thus, a strategy had to be defined to consider dialogue success. More specifically, it was considered that the dialogues were successful when the user obtained the information he requested.

All the annotations were stored in a database from which the values for the interaction parameters were automatically computed. For example, dialogue duration was computed from the timestamps, and the number of confirmation turns was computed using the information about the previous system turn. Table 5.1 sets out the interaction parameters that were used in the experiments, while Figure 5.2 shows the parameter computation for a sample dialogue.

Parameter	Description	Human annotation needed
Task success	Binary value which indicates whether the user obtained from the system the information he requested	Yes
Dialogue completion	Binary value which indicates whether the user waited until the end of the dialogue before hunging up	Yes
Dialogue duration	Duration of the dialogue in seconds	No
Number of user turns	Number of user turns in the dialogue	No
Average number of words per turn in the dialogue	Avg. number of words along all the user utterances in the dialogue	Yes
WER	Number of incorrectly recognized words divided by the total number of words uttered by the user in the dialogue	Yes
Average recognition confidence	Average confidence score of the words in all the speech recognition results in the dialogue	No
% correctly understood utterances	Percentage of correctly understood utterances with respect to the total number of utterances in the dialogue	Yes
Number of confirmation turns	Number of times the system asked for explicit confirmation during the dialogue	No

Table 5.1. *Interaction parameters employed*

DIALOGUE

S1> Welcome to the UAH system. How can I help you?

U1> I want information about a professor.

S2> Please say the name of the professor.

U2> Peter Smith.

S3> Did you say Peter Smint?

U3> No, Peter Smith.

S4> What information do you want about Professor Smith?

U4> I would like to know his tutoring timetable.

S5> Professor Smith's tutoring hours are on Mondays and Tuesdays from 10 to 12. Do you need any other information?

U5> No, thanks.

S6> Thank you for using the UAH system. Have a nice day!

PHASE 1 :
Automatic call recording and information registration

2006_04_07, 15:02:48 – 15:03:59, Male

	Timestamp	Previous system turn	Recognized sentence	Recognition confidence
U1	15:03:09	Initial_menu	I want information about professors	0.89
U2	15:03:20	Ask_prof_name	Peter Smint	0.4
U3	15:03:36	Conf_prof_name	No Peter Smith	0.9
U4	15:03:43	Prof_menu	I would like to know his tutoring timetable	0.85
U5	15:03:57	More_info	No thanks	1

PHASE 2 : Interaction parameters computation

2.1: Annotation by human expert

Task success	1
Dialogue completion	1
Number of words per turn	U1 – 6 U2 – 2 U3 – 3 U4 – 8 U5 - 2
Number of insertions, deletions and modification per turn	U1 – 1 deletion, 1 substitution U2 – 1 substitution
Correctness of the semantic interpretation	U1 – 1 U2 – 0 U3 – 1 U4 – 1 U5 – 1

2.2: Automatic computation

Dialogue duration	71
Number of user turns	5
Average recognition confidence	0.81
Number of confirmation turns	1
Average number of words/turn	4.2
WER	0.14
%correctly understood utterances	0.8

Figure 5.2. *Example of the computation of the interaction parameters for an UAH dialogue*

5.3.2. Quality judgments

The interaction with the UAH system starts with a welcome message in which the system introduces itself, and asks the user to visit a web page where he can complete a questionnaire with his opinion about the system performance. To be able to link the results of this test with the recordings of the user-system interaction, the user is provided with a dialogue identification number. This number is requested in the questionnaire along with the date he made the telephone call to the system and an approximate time for the start of the interaction.

The English translation of the questionnaire is as follows:

Q1. State on a scale from 1 to 5 your knowledge about new technologies for information access. (1 = "Low", 5 = "High")
Q2. State on a scale from 1 to 5 your previous experience using telephone-based dialogue systems. (1="Low", 5="High")
Q3. How many times have you used the UAH system before?

- I have not used it before.
- ..... times.

Q4. How well did the system understand you?

- Extremely bad.
- Bad.
- Fair.
- Good.
- Excellent.

Q5. How well did you understand the messages generated by the system?

- Extremely bad.
- Bad.
- Fair.
- Good.
- Excellent.

Q6. In your opinion the interaction was:

- Very slow.
- Slow.
- Adequate.
- Fast.
- Very fast.

Q7. Correcting the errors made by the system was:

- Extremely difficult.
- Difficult.
- Easy.
- Extremely easy.
- The system made no errors.

Q8. Was it easy for you to get the information that you requested?

- No, it was impossible.
- Yes, but with great difficulty.
- Yes, but with certain difficulties.
- Yes, it was easy.
- Yes, it was extremely easy.

Q9. Are you satisfied with the system performance?

- Not satisfied at all.
- Not very satisfied.
- Indifferent.
- Satisfied.
- Very satisfied.

Q10. Were you sure about what to say to the system at every moment?

- No, never.
- No, almost never.
- Sometimes.
- Yes, almost always.
- Yes, always.

Q11. Do you believe the system behaved similarly as a human would do?

- No, never.
- No, almost never.
- Sometimes.
- Yes, almost always.
- Yes, always.

The answers to each question were encoded and appropriately saved in the interactions database. All the answers excepting those corresponding to Q3 were assigned a numeric value between one and five (in the same order as they appear in the questionnaire). The values by default were: Q1=1, Q2=1, Q3=1, Q4=3, Q5=3, Q6=3, Q7=5, Q8=3, Q9=3, Q10=3, Q11=3. From the results of the test, the measures listed in Table 5.2 were extracted:

The first three measures listed in Table 5.2 are not quality judgments, but information about users. With the help of these questions, we intended to obtain an approximate idea of the users' background. However, as the UAH users were mainly students and professors of our Faculty, knowledge about new technologies for information access was high in almost all cases, as it is shown in Figure 5.3. Only 36% of our test participants were women.

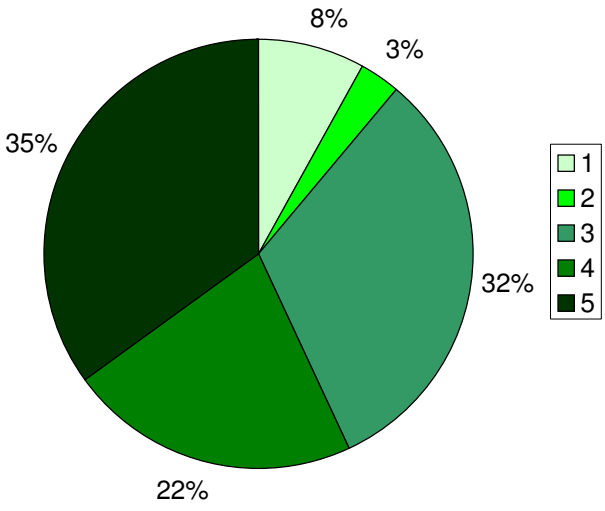


Figure 5.3. *Users' knowledge about new technologies for information access (1 = Low, 5 = High)*

Parameter	Question from which it is extracted
Knowledge about new technologies for information access	Q1
Knowledge about dialogue systems	Q2
Experience using the UAH system	Q3
Perceived extent to which UAH understands the user	Q4
Perceived extent to which the user understands UAH	Q5
Perceived interaction speed	Q6
Perceived presence of errors made by UAH	Q7
Perceived ease of UAH error correction	Q7
Perceived easy of obtaining the requested information	Q8
User satisfaction	Q9
Extent to which the user knew what to say at each moment of the interaction	Q10
Perceived human-like behaviour of the UAH system	Q11

Table 5.2. *Perceived quality and user profile parameters employed*

As our experiments were based on calls made by users who phoned the system on their own initiative, we think that the results obtained are very realistic, given that the interaction was based on a real need of the users. Besides, dialogues were more heterogeneous as they take place in different contexts. The disadvantage of this approach was that, although the users were encouraged to answer the questionnaires, some of them did not do it, and thus there were no quality judgments for all the recorded dialogues. Specifically, only 37 of the 85 dialogues have subjective measures along with the objective ones. Figure 5.4 shows the demographic data of the two types of users: those who answered the subjective test, and those who did not.

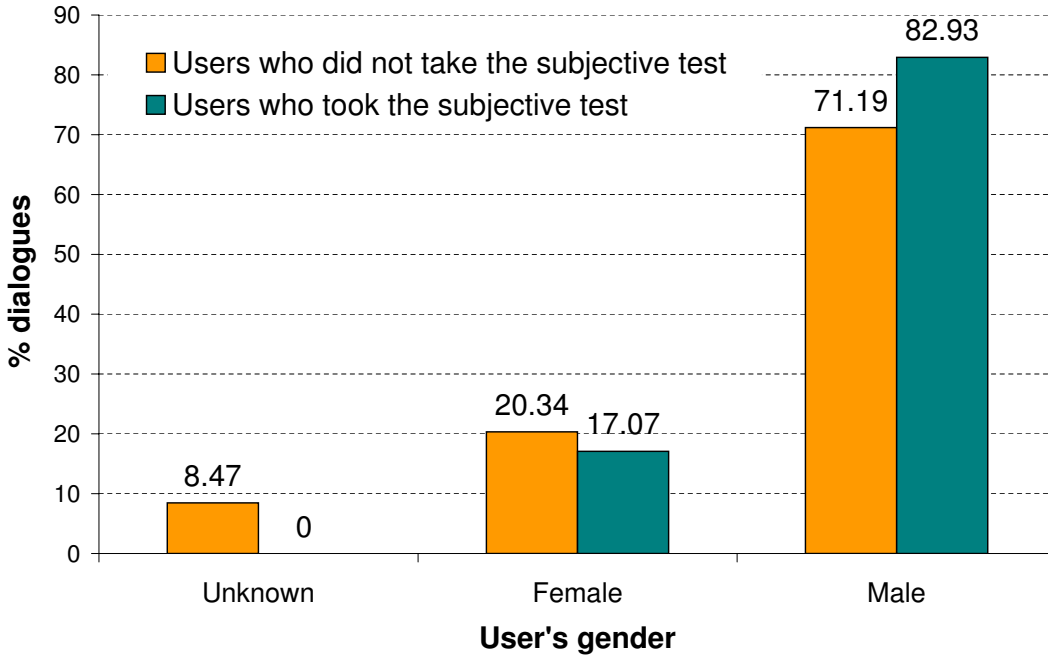


Figure 5.4. *Demographic data for the different user types*

As can be observed, from the dialogues that corresponded to users who did not fill in the questionnaire, 8.47% were annotated with an unknown gender of the speaker. This is because these users hung up after the first prompt of the system and said nothing in response. The first system prompt clearly stated that the user was about to talk to an automatic system, and that the call was going to be recorded for research purposes. Hence, we think that two plausible reasons why some users hung up before their first turn are that they did not feel confident in talking to a computer, and that they were not happy with having their interactions recorded.

The descriptive statistics of all the parameters regarding the type of users involved are shown in Table 5.3, where the minimum, maximum and range values of all the measures used in our study are indicated. Section 5.5.1 presents a detailed study of the differences in performance and perceived quality between the interactions of these two user groups.

Parameter	User type	Range	Min.	Max.	Avg.	Typ. Dev.	Variance
Knowledge about new technologies for information access	Subj. test	4	1	5	3.77	1.14	1.30
Knowledge about dialogue systems	Subj. test	4	1	5	3.23	1.28	1.65
Experience using the UAH system	Sub. test	9	1	10	2.80	3.20	10.22
Perceived extent to which UAH understands the user	Subj. test	4	1	5	3.69	1.25	1.57
Perceived extent to which the user understands UAH	Subj. test	2	3	5	4.37	0.69	0.48
Perceived interaction speed	Subj. test	3	1	4	2.71	0.62	0.39
Perceived presence of errors made by UAH	Subj. test	1	0	1	0.54	0.50	0.25
Perceived ease of UAH error correction	Subj. test	3	1	4	2.47	0.90	0.82
Perceived easy of obtaining the requested information	Subj. test	4	1	5	3.37	1.437	2.06
User satisfaction	Subj. test	4	1	5	3.63	1.09	1.18
Extent to which the user knew what it was expected from him at each point of the dialogue	Subj. test	3	2	5	4.29	0.893	0.798
Perceived human-like behaviour of the UAH system	Subj. test	4	1	5	3.57	1.04	1.08
Task success	Subj. test	1	0	1	0.77	0.43	0.18
	No subj. test	1	0	1	0.46	0.50	0.25
Dialogue completion	Subj. test	1	0	1	0.74	0.44	0.20
	No subj. test	1	0	1	0.36	0.48	0.23
Dialogue duration	Subj. test	153	21	174	96.66	37.06	1373.70
	No subj. test	297	0	297	90.14	64.65	4179.88
Number of user turns	Subj. test	9	1	10	5.34	2.26	5.11
	No subj. test	16	1	17	4.7	3.94	15.52
Avg. words per turn	Subj.test	3	1	4	1.81	0.69	0.48
	No subj. test	4.33	0	4.33	1.73	0.78	0.61
WER	Subj. test	0.67	0.00	0.67	0.19	0.18	0.03
	No subj. test	0.83	0	0.83	0.25	0.28	0.05
Avg. recognition confidence	Subj. test	0.16	0.82	0.98	0.93	0.04	0.002
	No subj. test	0.23	0.77	1	0.93	0.05	0.003
% correctly understood utterances	Subj. test	0.50	0.50	1	0.95	0.12	0.15
	No subj. test	0.74	0.33	1	0.89	0.19	0.04
Number of confirmation turns	Subj. test	2	0	2	0.80	0.63	0.40
	No subj. test	3	0	3	0.62	0.88	0.77

Table 5.3. *Descriptive statistics of the criteria used*

5.4 Statistical studies employed for evaluation

In order to find relevant relationships between the criteria used, all the variables were correlated, obtaining the absolute value of the *Pearson correlation coefficient*. However, the value of the correlation coefficient by itself was not

enough to obtain reliable results, as it was also necessary to know the probability of obtaining the results by chance. This was done by computing the significance (or *p-value*) of each correlation coefficient. If the significance level was very small (less than 0.05) then the correlation was significant and the two criteria were considered linearly related.

As most of the variables were inter-correlated, the effect that each criterion had on the significance of the relationships between the rest has been studied. It is possible that two criteria are correlated just because they are both affected by a third one. Thus, when eliminating the effect of this criterion, they would not be significantly correlated. To study the relationships in isolation, eliminating the effect of the rest of the criteria, the *partial correlation coefficients* were computed along with their significance.

The *Pearson correlation coefficient* is suitable for *scale* variables, whose values represent ordered categories with meaningful metrics, such as dialogue duration in seconds, so that distance comparisons between the values are appropriate. However, not only scale variables were used, but also *ordinal* and *dichotomous* variables (a classification can be found in Table 5.4). The values of the ordinal variables represent categories with an intrinsic rating, such as the perceived quality parameters described in Section 5.3.2. Dichotomous variables, such as "task success" or "dialogue completion", can only have two values (0 or 1 in our case). Thus, in order to obtain reliable results, contingency tables were built for the ordinal criteria. These Tables allow to study these variables and discover associations between them. To measure the strength of their relationships, the *Kendall's Tau-b* and the *Spearman's rho* coefficients were used. The interpretation of these coefficients is equivalent to that of the *Pearson coefficient*. However, as they are based on the ordinal properties of the data, their values and significances may not be the same.

Additionally, analyses of variance (ANOVA) were carried out. Essentially, ANOVA models try to describe a dependent variable as the result of the weighted sum of several factors. Specifically, one-way ANOVA was used, in which there is only one independent variable, and computed the *F coefficient*. When *F's* critical level is below 0.05, it is possible to discard the average equality and conclude that not all the poblational averages that are being compared are equal. *Eta square* was also obtained, which is an estimation of the degree to which each factor affects the dependent variable.

To obtain more information on which to base our interpretations, especially for the case of dichotomous variables, *Phi* and *Cramer's V* coefficients were also calculated, which allow to contrast the independence hypothesis in contingency tables.

Parameter	Type
Knowledge about new technologies for information access	Ordinal
Knowledge about dialogue systems	Ordinal
Experience using the UAH system	Ordinal
Perceived extent to which UAH understands the user	Ordinal
Perceived extent to which the user understands UAH	Ordinal
Perceived interaction speed	Ordinal
Perceived presence of errors made by UAH	Dichotomous
Perceived ease of UAH error correction	Ordinal
Perceived easy of obtaining the requested information	Ordinal
User satisfaction	Ordinal
Extent to which the user knew what it was expected from him at each point of the dialogue	Ordinal
Perceived human-like behaviour of the UAH system	Ordinal
Task success	Dichotomous
Dialogue completion	Dichotomous
Dialogue duration	Scale
Number of user turns	Scale
Avg. words per turn	Scale
WER	Scale
Avg. recognition confidence	Scale
% correctly understood utterances	Scale
Number of confirmation turns	Scale

Table 5.4. *Type of variables used for the statistical studies*

All the experiments were carried out using the SPSS 14 predictive analysis software[2]. For the experiments in which the aim was to obtain important relationships between all the evaluation criteria including both interaction parameters and quality judgments, the 37 dialogues in which the users answered the subjective test were used. For the experiments in which the possible reasons for the users to take the test or not were studied, both types of dialogues were used (85 in total).

[2]Statistical Product and Service Solutions - http://www.spss.com/

5.5 Evaluation results

This section presents a summary of the numeric results obtained from the statistical studies. Table 5.5 shows a summary of the results obtained with the partial correlations. For reasons of space not all the 21 partial correlations tables with their numeric values have been reported. Instead, there are only reported all the significant correlations found between all the tables, along with the number of control criteria for which they were significant (i.e. the number of partial correlation tables in which the relationship was significant). For each pair of criteria, Table 5.6 sets out the *Pearson correlation* coefficient and its significance level. Significance levels below 0.05 are marked in blue, those below 0.01 are marked in orange, and non-significant relations are left white.

As can be observed, there were no significant relations regardless of the control criteria used (i.e. none of them appeared in the 21 tables). In fact, the best case was achieved when the relationship between two criteria was shown to be significant when eliminating the effect of 17 of the 21 variables. This showed that all the variables were deeply related. Finally, in Table 5.7 there is a summary of the results for the *Tau-b* and *Rho* coefficients, only emphasizing the relations for which significance differs from those obtained in the Pearson correlation studies. In the following sections the main findings derived from these results are discussed and interpreted.

Criteria relationship		Partial correlation tables in which it was significant
Perc. ease of obtaining the requested information	Perc. extent to which UAH understands the user	17
Perc. human-like behaviour of the UAH system	Perc. extent to which the user understands UAH	17
Knowledge about dialogue systems	Knowledge about new technologies for information access	17
Dialogue duration	Number of user turns	16
Number of confirmation turns	Number of user turns	16
% correctly understood utt.	WER	16
Avg. recognition confidence	WER	16
Task success	Perc. ease of obtaining the requested information	16
Perc. ease of obtaining the requested information	User satisfaction	15
Perc. human-like behaviour of the UAH system	User satisfaction	15
Avg. recognition confidence	% correctly understood utt.	15
Task success	User satisfaction	15
Dialogue completion	Task success	14
Dialogue completion	User satisfaction	14
Perc. ease of UAH error correction	Perc. ease of obtaining the requested information	14
Perc. ease of UAH error correction	Perc. extent to which UAH understands the user	14
Task success	Perc. extent to which UAH understands the user	14
Perc. extent to which UAH understands the user	User satisfaction	14
WER	Avg. words per turn	14
Perc. ease of UAH error correction	User satisfaction	13
Perc. ease of obtaining the requested information	Dialogue completion	13
Perc. ease of obtaining the requested information	Perc. human-like behaviour of the UAH system	13
Dialogue completion	Perc. ease of UAH error correction	12
Dialogue completion	Perc. extent to which UAH understands the user	12
% correctly understood utt.	User satisfaction	12
Perc. ease of UAH error correction	Task success	12
Perc. human-like behaviour of the UAH system	Task success	12
Perc. human-like behaviour of the UAH system	Perc. extent to which UAH understands the user	12
Perc. ease of UAH error correction	Perc. human-like behaviour of the UAH system	11
Dialogue duration	Perc. ease of obtaining the requested information	10
Dialogue duration	Task success	10
Dialogue duration	Perc. extent to which UAH understands the user	10
Dialogue duration	User satisfaction	10
Task success	Number of user turns	10
User satisfaction	Perc. extent to which the user understands UAH	10
Dialogue completion	Dialogue duration	9
Number of user turns	User satisfaction	7
Dialogue completion	Perc. human-like behaviour of the UAH system	6
Number of confirmation turns	Dialogue duration	5
Perc. ease of obtaining the requested information	Perc. extent to which the user understands UAH	5
Number of user turns	Perc. ease of obtaining the requested information	4
User satisfaction	Avg. words per turn	4
Number of confirmation turns	Avg. recognition confidence	3
Dialogue duration	UAH usage	2
Dialogue completion	Number of user turns	1
Dialogue completion	UAH usage	1
Dialogue duration	Perc. ease of UAH error correction	1
Dialogue duration	Perc. human-like behaviour of the UAH system	1
Number of confirmation turns	UAH usage	1
Number of confirmation turns	WER	1
Number of user turns	WER	1
Perc. ease of obtaining the requested information	% correctly understood utt.	1
Perc. ease of obtaining the requested information	Avg. recognition confidence	1
Avg. recognition confidence	Task success	1
Avg. recognition confidence	UAH usage	1
Avg. recognition confidence	User sure	1
Task success	% correctly understood utt.	1
Perc. extent to which UAH understands the user	Number of user turns	1
Perc. extent to which UAH understands the user	% correctly understood utt.	1
UAH usage	WER	1
Knowledge about dialogue systems	Perc. ease of obtaining the requested information	1
Knowledge about dialogue systems	Perc. extent to which UAH understands the user	1
User satisfaction	WER	1

Table 5.5. *Significant partial correlations*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. NT knowledge	1																				
2. DS knowledge	0.619 0.000	1																			
3. UAH usage	0.116 0.506	0.226 0.191	1																		
4. UAH understands user	0.277 0.107	0.265 0.124	0.116 0.508	1																	
5. User understands UAH	0.149 0.394	0.034 0.846	0.288 0.093	0.343 0.044	1																
6. Interaction speed	0.113 0.519	-0.063 0.719	0.222 0.200	0.334 0.050	0.323 0.058	1															
7. Easy to correct errors	0.139 0.570	-0.190 0.435	0.095 0.699	0.809 0.000	0.267 0.269	0.387 0.101	1														
8. Perc. presence of errors	-0.034 0.848	-0.197 0.258	-0.131 0.453	-0.511 0.002	-0.258 0.135	-0.147 0.399		1													
9. Perc. easiness to get info.	0.197 0.257	0.128 0.464	0.074 0.672	0.784 0.000	0.628 0.000	0.353 0.038		-0.326 0.056	1												
10. User satisfaction	0.286 0.096	0.105 0.550	0.147 0.399	0.753 0.00	0.582 0.000	0.274 0.112		-0.318 0.063	0.844 0.000	1											
11. User sure	0.124 0.479	0.018 0.917	0.093 0.597	0.476 0.004	0.539 0.001	0.204 0.239	0.234 0.335	-0.419 0.012	0.488 0.003	0.385 0.022	1										
12. UAH human behaviour	0.238 0.168	0.208 0.230	0.284 0.098	0.684 0.000	0.764 0.000	0.443 0.008	0.601 0.006	-0.385 0.022	0.761 0.000	0.690 0.000	0.580 0.000	1									
13. Dialogue completion	0.113 0.518	0.003 0.987	0.212 0.222	0.485 0.003	0.321 0.060	0.152 0.382	0.623 0.004	-0.146 0.402	0.524 0.001	0.589 0.000	0.117 0.505	0.329 0.054	1								
14. Task success	0.253 0.143	0.098 0.575	0.052 0.768	0.742 0.000	0.498 0.002	0.301 0.079	0.623 0.004	-0.226 0.191	0.912 0.000	0.827 0.000	0.331 0.052	0.637 0.000	0.614 0.000	1							
15. Dialogue duration	-0.037 0.834	-0.072 0.682	0.104 0.554	0.433 0.009	0.155 0.375	0.215 0.215	0.421 0.073	-0.177 0.310	0.348 0.040	0.375 0.026	0.245 0.157	0.226 0.191	0.462 0.005	0.475 0.004	1						
16. # user turns	0.054 0.758	0.134 0.442	-0.173 0.320	0.340 0.046	-0.046 0.792	0.030 0.865	0.160 0.513	-0.193 0.266	0.249 0.149	0.257 0.137	0.183 0.293	0.027 0.878	0.354 0.037	0.328 0.054	0.744 0.000	1					
17. Recog. confidence	0.152 0.584	0.149 0.394	-0.039 0.822	0.110 0.528	0.114 0.516	-0.032 0.855	0.133 0.587	-0.059 0.736	0.092 0.599	-0.024 0.891	0.315 0.065	0.246 0.154	0.070 0.688	-0.028 0.874	-0.027 0.876	0.098 0.576	1				
18. WER	-0.168 0.336	-0.034 0.845	-0.279 0.104	-0.157 0.369	-0.254 0.141	-0.134 0.443	0.081 0.742	0.191 0.273	-0.093 0.595	0.014 0.936	-0.426 0.011	-0.228 0.187	-0.163 0.349	-0.017 0.925	0.040 0.821	0.244 0.158	-0.526 0.001	1			
19. Words per turn	0.154 0.377	0.015 0.931	-0.057 0.745	-0.103 0.555	-0.107 0.542	-0.150 0.389	0.242 0.318	0.215 0.215	0.036 0.840	0.213 0.219	-0.158 0.364	-0.110 0.528	0.198 0.253	0.168 0.333	0.027 0.880	0.096 0.585	-0.323 0.058	0.539 0.001	1		
20. % understood utt.	0.197 0.256	0.080 0.647	0.212 0.221	0.469 0.004	0.249 0.150	-0.009 0.959	0.312 0.194	-0.341 0.045	0.350 0.040	0.495 0.002	0.398 0.018	0.379 0.025	0.291 0.089	0.329 0.054	0.115 0.510	0.027 0.878	0.440 0.008	-0.507 0.002	-0.065 0.713	1	
21. # confirmation turns	0.057 0.744	0.094 0.591	-0.297 0.083	0.363 0.032	-0.027 0.878	0.000 1.000	0.227 0.350	-0.478 0.004	0.181 0.297	0.188 0.179	0.312 0.068	0.090 0.608	0.126 0.471	0.153 0.381	0.598 0.000	0.749 0.000	0.305 0.075	0.034 0.847	-0.142 0.416	0.150 0.390	1

Table 5.6. *Correlations between the criteria used*

Criterion 1	Criterion 2	*Pearson*	*Tau-b*	*Rho*
Perc. extent to which UAH understands the user	DS knowledge	0.265 0.124	0.276 0.057	0.336 0.049
Perc. extent to which UAH understands the user	Perc. interaction speed	0.334 0.050	0.268 0.077	0.299 0.081
Perc. extent to which UAH understands the user	Dialogue completion	0.485 0.003	0.390 0.013	0.426 0.011
Perc. extent to which UAH understands the user	Dialogue duration	0.433 0.009	0.209 0.111	0.278 0.105
Perc. extent to which UAH understands the user	Number of user turns	0.340 0.046	0.157 0.255	0.197 0.257
Perc. extent to which UAH understands the user	Number of confirmation turns	0.363 0.032	0.291 0.054	0.335 0.049
Perc. extent to which the user understands UAH	Task success	0.498 0.002	0.408 0.013	0.424 0.011
Perc. human-like behaviour of the UAH system	Perc. interaction speed	0.443 0.008	0.355 0.019	0.389 0.021
Perc. human-like behaviour of the UAH system	Perc. ease of UAH error correction	0.601 0.006	0.474 0.018	0.523 0.022
Dialogue completion	Perc. ease of UAH error correction	0.623 0.004	0.559 0.011	0.602 0.006
Task success	Perc. ease of UAH error correction	0.623 0.004	0.559 0.011	0.602 0.006
Perc. easy of obtaining the required information	Perc. presence of errors made by UAH	-0.326 0.056	-0.337 0.033	-0.365 0.031
User sure	Perc. presence of errors made by UAH	-0.419 0.012	-0.429 0.008	-0.454 0.006
User sure	User satisfaction	0.385 0.022	0.291 0.054	0.316 0.064
Dialogue duration	User satisfaction	0.375 0.026	0.245 0.065	0.310 0.070
% correctly understood utt.	User satisfaction	0.495 0.002	0.223 0.151	0,248 0.151
Perc. easy of obtaining the requested information	Dialogue completion	0.524 0.001	0.384 0.015	0.416 0.013
Dialogue duration	Dialogue completion	0.462 0.005	0.350 0.014	0.421 0.012
Number of user turns	Dialogue completion	0.354 0.037	0.274 0.068	0.313 0.067
Perc. easy of obtaining the requested information	Dialogue duration	0.348 0.040	0.151 0.253	0.225 0.194
Dialogue duration	Task success	0.475 0.004	0.362 0.011	0.435 0.009
Dialogue completion	Number of user turns	0.354 0.037	0.274 0.068	0.313 0.067
User sure	WER	-0.426 0.011	-0.337 0.017	-0.388 0.021
Perc. easy of obtaining the requested information	% correctly understood utt.	0.350 0.040	0.244 0.113	0.262 0.129

Table 5.7. *Significance variations between Pearson, Chramer's Tau-b and Spearman's Rho*

5.5.1. Impact of the interaction performance on the user decision to answer the subjective test

As was described in Section 5.3.2, not all the users answered the subjective test from which the perceived quality criteria were computed. In order to study if there were some interaction parameters that influenced the users' decision to answer the test, we introduced a dichotomous variable indicating whether the user answered the test or not, and carried out Pearson correlation and ANOVA studies to find its relationship with the interaction parameters. Table 5.8 shows the results obtained.

Relationship	ANOVA F (Sig)	Eta square	Pearson (Sig)
Task success	7.156(0.009)	0.079	0.282(0.009)
Dialogue completion	7.775(0.007)	0.086	0.293(0.007)
Dialogue duration	0.245 (0.622)	0.003	0.054 (0.622)
Number of user turns	0.729 (0.396)	0.009	0.093 (0.396)
Avg. recognition confidence	0.122 (0.728)	0.001	-0.159 (0.150)
WER	2.107 (0.150)	0.025	0.010 (0.927)
Avg. words per turn	0.008 (0.927)	0.000	0.038 (0.728)
% correctly understood utt.	3.759 (0.056)	0.043	0.208 (0.56)
Number of confirmation turns	0.592 (0.447)	0.18	0.133 (0.447)

Table 5.8. *Significance of the relationship between "The user taking the subjective test" and the interaction parameters*

The only relations that were shown to be significant for the "user taking the subjective test" were with the "dialogue completion" and the "task success" metrics. These are two criteria that were also very significantly correlated with each other, with an *ANOVA F* of 180.159, and a 0.000 significance. *Eta square* was 0.685, and as both are dichotomous variables, *Phi* and *Cramer's V* were also calculated, obtaining for both coefficients a value of 0.827 and a 0.000 approximate significance.

One conclusion to be derived from these results is that the users carried out the subjective test mainly when they succeeded in getting the information they wanted. The fact that the successful dialogues were related to dialogue completion might be because unsuccessful dialogues were usually prematurely finished by the user.

To check whether the interaction parameters that affect task success are the same for all the user groups, additional ANOVA studies were carried out, which yielded the results shown in Table 5.9.

Relationship	User group	F	Sig
Dialogue completion - Task success	Users who did not take the subjective test	93.312	0.000
	Users that took the subjective test	19.951	0.000
	All users	180.159	0.000
Dialogue duration - Task success	Users who did not take the subjective test	17.814	0.000
	Users that took the subjective test	9.638	0.004
	All users	21.532	0.000
Number of user turns - Task success	Users who did not take the subjective test	13.025	0.001
	Users that took the subjective test	3.977	0.054
	All users	16.231	0.000
Avg. recognition confidence - Task success	Users who did not take the subjective test	0.105	0.748
	Users that took the subjective test	0.026	0.874
	All users	0.789	0.377
WER - Task success	Users who did not take the subjective test	0.171	0.681
	Users that took the subjective test	0.009	0.925
	All users	0.292	0.590
Avg. words per turn - Task success	Users who did not take the subjective test	12.787	0.001
	Users who took the subjective test	0.964	0.333
	All users	15.452	0.000
% correctly understood utt. - Task success	Users who did not take the subjective test	5.891	0.019
	Users who took the subjective test	3.992	0.054
	All users	12.539	0.001
Number of confirmation turns	Users who did not take the subjective test	0.528	0.471
	Users who took the subjective test	0.789	0.381
	All users	0.963	0.334

Table 5.9. *ANOVA table for task success and the rest of the interaction parameters regarding the different user groups*

As can be observed in the table, the only differences related to task success appeared for its relationships with the number of user turns, the percentage of correctly understood words per turn, and the number of words per turn. The three relationships were significant for the users who did not answer the test, but not for those who answered it, although the first two cases can be considered as almost significant at the 0.05 level. This change might be due to the degree of cooperation of the different types of user. For example, the users who did not answer the test and had unsuccessful dialogues, hung up immediately: 70.37% of the times before the fourth user turn. However, the users who answered the subjective test were more patient and tried to overcome the interaction problems even when in the end they could not obtain the information that they were asking for.

The main difference detected between both user groups was in the relationship between the number of words per turn and task success. For the users who did not answer the test, F had a value of 12.787 and it was significant below the 0.01 level, whereas for those who answered the test, F was 0.964 and it was not significant. This was probably because the distribution of the number of words per turn for the unsuccessful and successful dialogues was more balanced in the case of the users who answered the subjective test. For them, successful and unsuccessful dialogues had a similar number of words per turn. However, the users who did not answer the test employed no more than an average of one word per turn in their unsuccessful dialogues, and more than two turns in the successful ones. Thus, an average of words per turn less or equal to one was an indicator of dialogue failure in the case of users who did not answer the subjective test.

5.5.2. Criteria with highest impact on user satisfaction and task success

Table 5.10 shows the two highest correlation values with user satisfaction, which were obtained in all the statistical studies for the criteria "ease of obtaining information" and "task success". Thus, as expected, a user was highly satisfied when he found it easy to get the information he wanted. However, it is remarkable that the way of gathering information had the same order of significance with user satisfaction as with the final obtaining of the information. In (Möller, 2005), user satisfaction was also correlated

with the fact that the user finally obtained the information he was looking for. However, Möller's indicator of ease of communication (which he classified as a comfort factor) did not provide a significant contribution to the overall user satisfaction. This might suggest that ease of interaction is more important for users who have a real need to obtain the information from the system compared with those for whom the interaction is carried out by following predefined scenarios.

Relationship	*Pearson* (sig)	*Tau-b* (sig)	*Rho* (sig)	*ANOVA F* (sig)
Perceived easy of obtaining the requested information and User satisfaction	0.844 (0.000)	0.750 (0.000)	0.814 (0.000)	31.071 (0.000)
Task success and User satisfaction	0.827 (0.000)	0.732 (0.000)	0.787 (0.000)	33.140 (0.000)

Table 5.10. *Statistical significance of the most important relationships with "user satisfaction"*

In addition, the item of the subjective questionnaire from which the measure "perceived ease of use" is computed, implicitly takes into account the perceived success of the dialogue. Specifically, the answers to question Q8 in the questionnaire (Section 5.3.2) ranged from "No, it was impossible to get the information" to "Yes, it was very easy to get the information". Thus, there were two different task success measures: an interaction parameter that indicated whether the user was able to get the information that he was looking for, and another that indicated perceived task success. This second measure was extracted from the "ease of obtaining information" parameter by assigning 0 (unsuccessful) to the answer "No, it was impossible" and 1 (success) to the rest.

Contingency tables showed that both task success measures had the same value for all the dialogues. Hence, in our experiments task success was only considered as an interaction parameter. Previous studies such as (Rajman et al., 2004) found that as the users in laboratory tests are not given the possibility to contrast the information provided by the dialogue system, they trust the system responses. For example, they do not check whether the information is correct or useful. Thus, they consider the fact of obtaining a piece of information from the system equivalent to obtaining a correct result.

The authors studied this behaviour by employing laboratory test users who could not discern whether the information about restaurants, menus and prices provided by a dialogue system was correct. In our experiments, the UAH users were provided with real academic information. As they had a real need for this information, they could contrast it and know whether it was accurate or not. Thus, among the unsuccessful dialogues (both from the interaction parameter and the quality perception points of view) there were cases where the system provided information to the user but it was not what they desired, as is shown by the fact that some complete dialogues were unsuccessful. It is a benefit of test fields to allow this separation between the quality of the interaction and the quality of the results.

Within interaction parameters, there is a remarkably high correlation between dialogue completion and task success. As shown in Figure 5.5, although users could hang up when they received the desired information, without waiting for the system to ask if they needed any other information, if the dialogue was successful, they usually waited until the end. Although the percentage of complete and successful dialogues was higher for more collaborative users (i.e. those who answered the questionnaire), both the users that took the subjective test and those who did not take it were patient enough to wait until the end of the dialogue when it was successful.

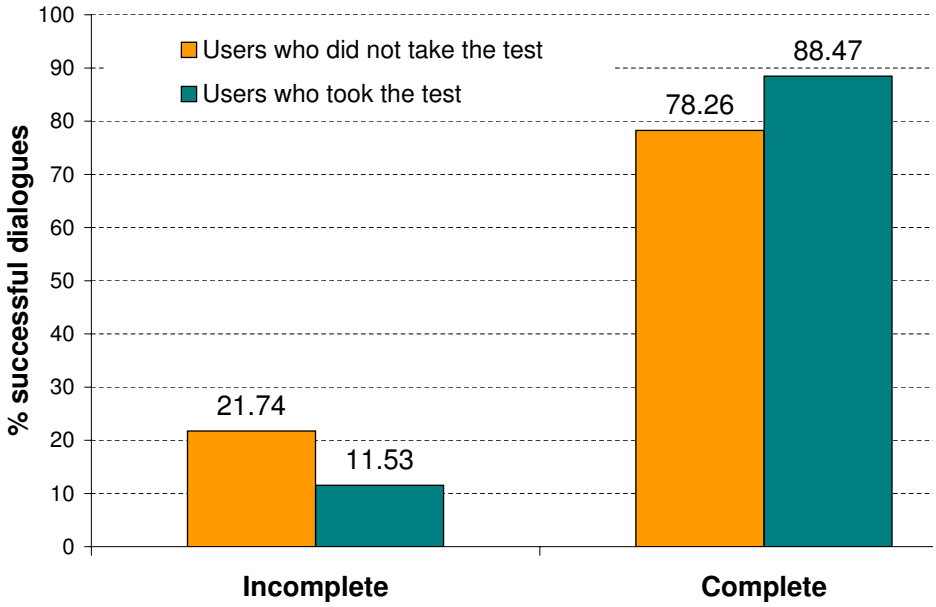


Figure 5.5. *Percentage of successful dialogues which are also complete regarding the different user groups*

This differs from findings of other authors. For example, Turunen et al. (2006) reported that there were highly significant differences on how the interaction was finished in field and laboratory tests carried out with the Stopman system. In the laboratory tests, 65% of the users employed an explicit request to end the call (e.g. "thank you and goodbye"). On the contrary, in the field tests less than 10% of users waited to the end of the call before hanging up. The number of dialogues in which the users waited until the end of the interaction (i.e. the number of complete dialogues) in our field study is more than 50% higher than in that of Turunen et al. (2006).

Rajman et al. (2004) discuss that a positive attitude of users towards a system does not only depend on its behaviour, but also on the "technophile" or "technophobe" attitude of the users, although they did not control these parameters in their experimentation. In our experiments, 57% of the users rated their knowledge about new technologies for accessing information above 3 in a 1-5 scale, where 1 represented "low" and 5 "high". Thus, the collaborative nature of our users could be a result of their possible technophile disposition.

Another criterion which is highly correlated with task success and user satisfaction is the perceived ease of error correction. However, the perceived presence of errors is not significantly correlated with any of these criteria. This is probably because although in 48.19% of the successful dialogues the users detected errors, in most cases they managed to circumvent them and obtain the information they were looking for. Specifically, as shown in Figure 5.6, the 69.23% of the users found it "easy" or "very easy" to correct errors in the successful dialogues. However, in the non-successful ones, 83.33% of the users found it "difficult" or "very difficult" to correct the errors.

In (Möller, 2005), the users' opinion about whether misunderstandings could be easily clarified, which was classified as a contributing factor to dialogue smoothness, was not a good predictor for user satisfaction. Additionally, the author found that user satisfaction could not be fully predicted by task success, and argued that this result could be because of the unrealistic situation of the laboratory experimentation employed. It has been corroborated this finding in our field study, as the subjective user tests could not be replaced by the interaction parameters employed without losing information.

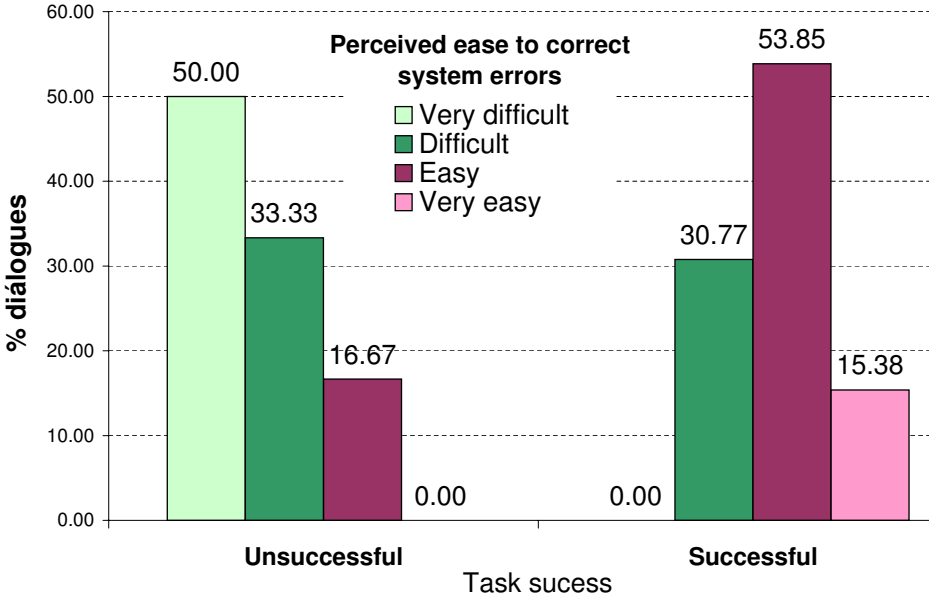


Figure 5.6. *Task success vs. Perceived ease of error correction*

5.5.3. Criteria with highest number of significant relations

The criterion that showed the largest number of significant correlations was the "perceived extent to which UAH understands the user". On the one hand, it was highly correlated with other quality judgments, like the degree to which the user understands the system, the perceived ease of error correction, the perceived ease of obtaining information, user satisfaction, perceived presence of errors (negative correlation in this case), and the perceived human-like behaviour of the system. Besides, as can be observed in Table 5.6, in most of these relations the significance was highest. On the other hand, this perceived quality criterion was highly correlated with interaction parameters such as completion of the dialogue, task success, dialogue duration or percentage of correctly understood utterances per dialogue.

The most significant relationships between this quality perception and other parameters were with task success and user satisfaction. Perceived system understanding, listed by Möller (2005) as an indicator of speech input quality, was very significantly correlated with user satisfaction in his study.

It is also interesting that the extent to which the user felt that the UAH system understood him was not correlated with the interaction parameters that measure the performance of the speech recognizer, such as WER or confidence scores. However, the percentage of correctly understood utterances was correlated with a significance below 0.01, which indicates that from the user's point of view, speech recognition errors were not important as long as the semantic interpretations were correct and thus these errors were invisible to him. This is reflected in that the perceived presence of errors was related to the percentage of correctly understood utterances and the number of confirmation prompts, but not to WER. However, perceived ease of error correction was not significantly correlated with any of these measures. Both the perceived presence of errors and the perceived ease of correcting them were very highly correlated with the perception that the UAH system understood the user. The perceived presence of errors also negatively affected the user's confidence about what to say next during the interaction.

5.5.4. Impact of user's knowledge and experience

It is noteworthy that the user's knowledge about dialogue systems and new technologies for accessing information were the criteria with the lowest correlation factors with all the others. However, they were significantly correlated with each other. Thus, in our case the knowledge of the user about new technologies for information access was not determinant on the results of the interaction, not in objective terms (e.g. duration, success), nor in perceived terms (e.g. perceived speed, user satisfaction). This may be because the great majority of users had a rather high level of technical knowledge. It is possible that in experiments with other dialogue systems, where users may have more varied backgrounds, these appear to be important criteria.

The previous experience of the user employing the system ("UAH usage") was not correlated with any of the other variables in any statistical study. However, the sign of the correlation parameters indicated that experienced users perceived fewer errors, needed fewer turns to get the information, provoked fewer recognition errors and required fewer confirmation turns.

The fact that previous UAH usage was not significantly correlated with other factors, such as task success or interaction speed, differs from results found in the literature. For example, Turunen et al. (2006) stated

that previous experience in using a system is a very important factor that can help to predict the success and smoothness of the dialogue. Similarly, Park et al. (2007) found that the performance of laboratory test users who had previously employed a system in very strictly predefined interactions was better than for those who had not employed it before. Other authors have studied the effect of user experience on quality judgments. For example Sturm et al. (2005) indicate that a previous prolonged use of the system helps to obtain substantial improvements in quality judgments, such as "ease of use" and "user satisfaction".

We believe that the impact of the user's experience is closely related to the type of evaluation carried out. In laboratory tests users are generally trained on how to employ the system, or at least are informed about how to interact with it. In field studies users commonly employ the system without any previous training, and this is why they are less prone to employ characteristics such as help requests (Turunen et al., 2006), of which they are sometimes not aware. However, these characteristics can be very useful to make interaction easier and to recover from error situations. On the other hand, in some particular areas of study, for example spoken dialogue systems for health applications, it has been argued that, contrary to what the previously commented studies suggest, an increasingly richer previous experience using the system does not always imply better performance and perceived quality results. For example (Bickmore and Giorgino, 2006) report that individuals who intermittently use health dialogue systems on the telephone, compared to those who use them frequently and those who hardly use them at all, obtain the highest satisfaction levels and the best outcomes in terms of the perceived benefits. However, as discussed by Farzanfar et al. (2004), this can be due to the stress that some users experience if they feel monitored.

5.5.5. Impact of dialogue management initiative

To study the impact of the initiative used for dialogue management, the computations discussed above were repeated, but distinguishing between dialogues with system-directed initiative and dialogues with mixed-initiative. The differences between both initiatives are reported in Table 5.11, where significant correlations are marked with 'Y' (yes) and non-significant with 'N' (no).

Criterion 1	Criterion 2	Mixed initiative	System-directed initiative
Perc. extent to which the user understands UAH	Perc. extent to which UAH understands the user	N	Y
Perc. interaction speed	Perc. extent to which UAH understands the user	Y	N
Perc. presence of errors made by UAH	Knowledge about dialogue systems	Y	N
Perc. presence of errors made by UAH	Perc. extent to which UAH understands the user	N	Y
User confidence about what to do next	Perc. presence of errors made by UAH	N	Y
User confidence about what to do next	Perc. ease of obtaining the requested information	N	Y
User confidence about what to do next	User satisfaction	N	Y
Perc. human-like behaviour of the UAH system	Perc. presence of errors made by UAH	N	Y
Perc. human-like behaviour of the UAH system	Perc. ease of obtaining the requested information	N	Y
Perc. human-like behaviour of the UAH system	User satisfaction	N	Y
Perc. human-like behaviour of the UAH system	User confidence about what to do next	N	Y
WER	Perc. presence of errors made by UAH	N	Y
Task success	User confidence about what to do next	N	Y
Task success	Perc. human-like behaviour of the UAH system	N	Y
Task success	Dialogue completion	Y	N
Dialogue duration	Dialogue completion	N	Y
Dialogue duration	Perc. ease of obtaining the requested information	Y	N
Dialogue duration	User satisfaction	Y	N
Dialogue duration	Task success	Y	N
Number of user turns	Task success	Y	N
Number of user turns	User satisfaction	Y	N
Number of user turns	Perc. ease of obtaining the requested information	Y	N
Dialogue completion	Perc. ease of obtaining the requested information	N	Y
Avg. recognition confidence	User satisfaction	Y	N
WER	User confidence about what to do next	N	Y
WER	Dialogue completion	N	Y
WER	Number of user turns	Y	N
% correctly understood utt.	Perc. ease of obtaining the requested information	N	Y
% correctly understood utt.	User satisfaction	N	Y
% correctly understood utt.	User confidence about what to do next	N	Y
% correctly understood utt.	Perc. human-like behaviour of the UAH system	N	Y
% correctly understood utt.	Dialogue completion	N	Y
% correctly understood utt.	Task success	N	Y
% correctly understood utt.	Avg. recognition confidence	N	Y
Number of confirmation turns	Perc. presence of errors made by UAH	N	Y
Number of confirmation turns	Dialogue duration	N	Y
Number of confirmation turns	Number of user turns	N	Y
Number of confirmation turns	Avg. recognition confidence	N	Y

Table 5.11. *Criteria that were significantly correlated with one initiative type but not with the other*

It was found that task success was approximately the same for both dialogue management initiatives. This differs from the results that can be found in the literature[3], where a more flexible initiative led to considerably higher task success rates. In our experiments success was higher for mixed initiative, but the difference between both was practically negligible (77.77% of the mixed-initiative dialogues and 76.92% of the system-directed ones were successful).

However, it was found that task success was related to different factors in each initiative. For example, in mixed-initiative dialogues the user's confidence about what to do next in the dialogue was not correlated with task success, user satisfaction or perceived ease of obtaining information. On the contrary, task success had a significant correlation with user confidence in system-directed dialogues. Probably this is because the user was less constrained in the mixed-initiative interactions, and hence he did not know exactly what he could say (Figure 5.7). This effect did not result in bad interaction results, as task success was not reduced in the case of mixed-initiative interactions.

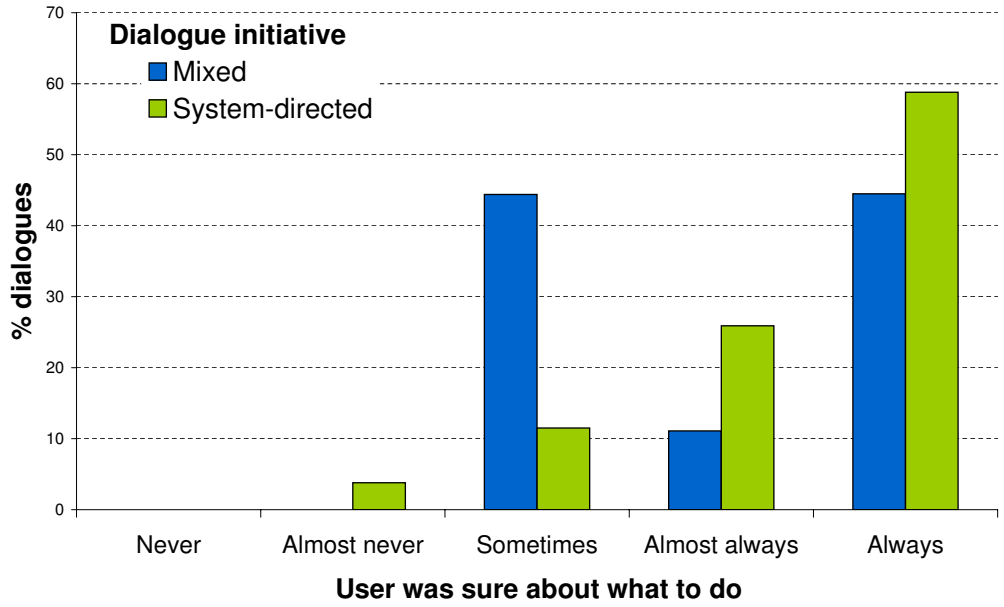


Figure 5.7. *Dialogue initiative influence on user confidence*

[3]A comprehensive summary can be found in (Möller, 2005)

Correlations of the perceived ease of obtaining information were also very different in the two cases. In the system-directed case it was related to the completion of the dialogue, the number of correctly understood utterances and the opinion that the user had about the human-like behaviour of the system. On the contrary, for mixed-initiative dialogues the perceived ease was not correlated with these measures, but with duration interaction parameters such as dialogue duration or number of user turns. The same happened with satisfaction (judgment) and task success (interaction parameter), which appeared to be highly correlated with duration measures in mixed-initiative interactions, but not in system-directed dialogues. The duration of these dialogues was significantly correlated with user satisfaction, whereas in restricted interaction systems this was not considered so important by users. Besides, as can be observed in Figure 5.8, the average duration of the dialogues was shorter when the interaction was more flexible (mixed-initiative instead of system-directed).

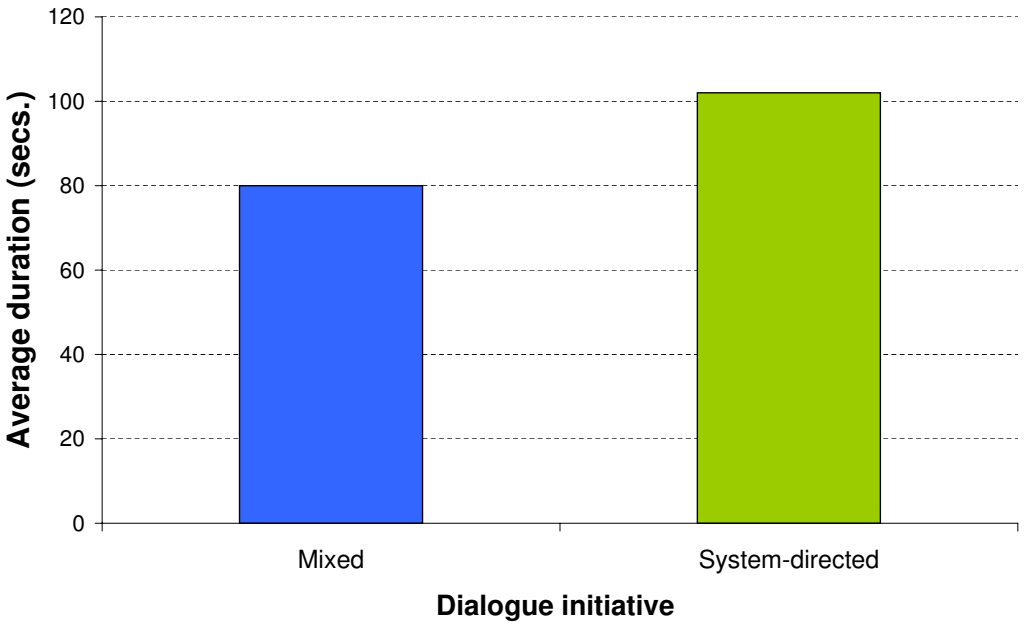


Figure 5.8. *Dialogue duration for each dialogue management intiative*

Additionally, the perceived presence of errors was related in mixed-initiative dialogues to the user's knowledge of dialogue systems. This was not the case for system-directed initiative. Besides, it was not correlated with other measures such as user confidence, WER or number of confirmation turns, which were important factors in the system-directed dialogues.

Studies based on laboratory tests like, for example, Rajman et al. (2004) could not find clear quality perception variations with respect to predominance of system or user-driven dialogue management initiatives. Besides, some laboratory tests like the one conducted for the BoRIS system in (Möller, 2005) could not find any significant relationship between the initiative experienced by users and other interaction parameters. However, our results show that the significance of the relationships between the different evaluation criteria, including both interaction parameters and quality judgments, vary depending on the initiative used for dialogue management.

5.6 Conclusions

In this chapter it has been presented a study of the relationships between several de-facto standard criteria for the evaluation of a telephone-based spoken dialogue system. Our experimental results are based on a field study using real interactions recorded from users who spontaneously telephoned the system to obtain information, without being recruited to do this.

To carry out our study both interaction parameters (or objective measures) and quality judgments (subjective measures) have been calculated by employing a corpus of real system-user interactions. Specifically, the quantitative criteria employed were: dialogue duration, dialogue completeness, task success, number of user turns, average recognition confidence, average WER, percentage of correctly understood utterances and number of confirmation turns. The qualitative measures were extracted from questionnaires that the users could optionally fill in. The criteria employed were: the extent to which the user felt correctly understood by the system, the extent to which the user understood the system messages, the perceived interaction speed, the perceived ease of error correction, the perceived presence of errors, the extent to which the user was sure about what he should do in every moment of the interaction, the extent to which the user believed the system's behaviour was human-like, and the level of user satisfaction with the interaction. Ad-

ditionally information about users was also taken into account, namely: user knowledge about new technologies for information access, user knowledge about spoken dialogue systems, and number of times the user had already used the system.

Several statistical studies were developed from which significant relations between all the criteria were extracted. This approach has not been sufficiently exploited in the literature, and some noteworthy empirical findings have been highlighted. Our empirical evidence shows that task success, perceived ease of obtaining information and perceived extent to which the system understands the user are very closely correlated with user satisfaction. These results suggest that obtaining the required information does not completely explain user satisfaction, as in some cases users judged successful dialogues as not satisfying because they found it difficult to obtain the information they are looking for. This is one of the implications derived from the usage of field tests, in which users are very concerned not only with obtaining the information they were looking for, but also with doing it easily. Furthermore, the relationship between the perceived ease of obtaining information and other criteria varies remarkably with the dialogue management initiative. Our experimental results show that, in the system-directed dialogues, perceived ease was related to the good functioning of the understanding module. On the contrary, in mixed-initiative dialogues, both user satisfaction and the perceived ease of obtaining the information seemed to be related to duration metrics. This had a strong implication in the quality judgments, as task success was highly correlated with user satisfaction in both initiatives. Thus, our results suggest that the prediction of user satisfaction also depends on the dialogue management initiative used. In the mixed-initiative dialogues it seemed to be more directly related to objective measures, such as dialogue duration. However, in more restricted dialogues, subjective measures such as the perceived extent to which the user feels that he is understood by the system, had a bigger impact. This is an important result that could indicate a need to tailor evaluation procedures to the type of interactions being analysed.

Additionally, we have studied the reasons that made some users answer the optional subjective test from which the quality judgments were obtained. It was found that it could be explained mainly in terms of dialogue completion and task success. Thus, the experiments carried out by including

the users' perceptions about the quality of the system, corresponded mainly to successful dialogues, in which the users obtained the information they were looking for. This could be one of the reasons why it was found that these users were very cooperative, which yielded high dialogue completion rates rarely reported in previous field test studies. Besides, contrary to what generally happens in laboratory studies, these measures consider that even when the user obtains information from the system, the dialogue cannot be considered successful if the provided information is not correct. Finally, no evidence was encountered of the effect of the users' previous experience employing the system on system performance or task success.

Hemos llegado al fin y yo inauguro
triste mi paz: la obra está completa.
Jorge Guillén, Obra completa

6 Conclusions and future work

6.1 Summary of contributions

The Thesis has introduced novel contributions directed towards the development of adaptive and portable spoken dialogue systems. Specifically, the Thesis presents approaches to build systems that recognize the users' emotions and adapt to their expectations and needs by detecting them in field evaluation studies. Additionally, portability is fostered by adapting the resources available in a language for the cost-effective development of speech-based systems in other languages.

A **complete spoken dialogue system, known as UAH, was developed** to evaluate the contributions of the Thesis. As described in Chapter 2, it provides academic information on the Dept. of Languages and Computer Systems. The system follows an architecture that is made up of five modules: a speech recognizer, an automatic grammar generation module, a dialogue manager, a database access module and an oral response generator. Given that the information provided by the UAH dialogue system is continually changing, a method was introduced to keep the recognition grammars updated with the last changes in the databases without introducing a delay in the interaction. The technique proposed is called **Triggered Grammar rules Creation (TGC)** and was implemented into the **Grammar Automatic Generation (GAG) module** of the UAH system. The system uses **different confirmation techniques** (explicit and implicit) as well as **several dialogue managing initiatives** (system-directed and mixed), in order to study their benefits both in terms of performance and perceived usability. The dialogue manager was developed using **VoiceXML documents that**

are dynamically generated according to the interaction context, which is also employed to tailor the system output to the user's needs.

The system was made public in June 2005. **From the interactions of the users with the system, an evaluation corpus has been generated and semi-automatically annotated**. The users' utterances were recorded in WAV format along with information about the recording starting time, the previous system turn, and the speech recognition result (including confidence scores). This information was stored in a database along with the dialogue starting and ending times. Nine parameters were automatically computed from this information and they were posteriorly employed to evaluate the system. Some of them were obtained automatically (e.g. dialogue duration), while others required manual annotation (e.g. dialogue success). Following this methodology an annotated corpus of **85 dialogues (422 user turns)** was obtained from the interactions of a year of user calls to UAH dialogue system. The users were invited to answer a questionnaire in which they could give their personal opinion about different aspects of their interaction with the system. From these opinions, a total of 12 quality criteria were obtained for each utterance including parameters such as the user satisfaction or the perceived interaction speed. Additionally, the corpus was annotated twice by nine non-expert annotators who classified each utterance in one of the following emotional categories: *neutral*, *doubtful*, *angry* and *bored*.

Chapter 3 has described the contributions of the Thesis which are focused on emotion recognition. Firstly, a **state of the art** on the main approaches used in the literature has been presented. It shows that research on emotion recognition has been mainly centred on how to apply different machine learning algorithms to differentiate between emotions, and less effort has been directed to determine the information in which the learning process should be based on. The most widespread approach consists of employing multimodal (audiovisual) information obtained from acted emotions, in order to retain strict control over the collected data.

The main contribution of the research carried out in the Thesis is the inclusion of contextual information for the **recognition of real emotions in spoken dialogue systems**. This has been a very challenging task, firstly because it had to be based on only one input modality, and secondly because natural emotions are expressed very subtly and usually the emotional categories encountered are very unbalanced, i.e. there is one predominant

emotion, which generally corresponds to the "neutral" state. This raises some problems that, to our knowledge, had not been fully addressed in literature before.

One of the main difficulties was to obtain reliable annotations of the emotional corpora considering that the traditional inter-annotator agreement measures are deeply affected by the skewning of the corpora obtained. In Chapter 3 we have presented a **detailed discussion on how to reliably calculate and interpret kappa coefficients** with corpora of real emotions. The proposed methodology has been **evaluated using the UAH emotional corpus and the significance of the results has been statistically computed**. The results of this evaluation show that, on the one hand, **our proposal makes possible to obtain annotator agreement values that are closer to the maximum attainable. It also allows non-expert annotators to detect more non-neutral emotions, and the annotation result is less affected by differences among the annotators**. On the other hand, **automatic emotion recognition obtains values that can be more than 40% better than those obtained with the approaches described in the literature**, which are based on the acoustic features without taking into account the proposed contextual information. The fact that all the proposals have been tested with a corpus of real emotions makes **the results directly applicable to practical operations**, as all **the algorithms proposed can be used dynamically during the execution of the dialogue systems**.

Chapter 4 has presented the work carried out during a three months stay in the Laboratory of Computer Speech Processing in the Technical University of Liberec under the supervision of Prof. Jan Nouza. As stated in the chapter, the application of the resources that are already available for a specific language to recognize a different language facilitates the development of new speech recognizers, specially in the case of minority languages and dialects. The approaches that can be found in literature usually demand laborious phonetic and linguistic studies of the languages involved. **A new method has been proposed to adapt a speech recognizer to other languages in a efficient way**. Our proposal has been evaluated using the MyVoice system, which was developed in the Technical University of Liberec for handicapped Czech users, and whose **translation to the Spanish language** is presented in the Thesis. It has been empirically demonstrated

that the proposed cross-lingual adaptation can be performed in a relatively short time by carrying out an expert-driven correspondence between both languages' phonetic alphabets. Experimental results show that for a task involving a vocabulary of 432 commands, 95.6% performance can be attained for Spanish and 97.5% for Slovak after adapting a Czech recognizer. For vocabularies up to 149k words, 72.9% and 77.4% accuracy rates were obtained for Spanish and Slovak respectively. The results are very promising given that they show that that portability of speech recognizers can be ensured fairly simply and that **the approach can achieve good results with similar languages such as Czech and Slovak, and also for phonetically different languages such as Czech (Slavic) and Spanish (Italic)**.

In Chapter 5 a **field evaluation of the UAH system** has been presented. Typically, in the literature, evaluation is carried out under restricted laboratory conditions, in which users are recruited to test the systems following a predefined list of scenarios. The problem with this method is that the scenarios may differ from the tasks that a user would have selected in a real interaction. In the field study that has been carried out in the Thesis, the users interacted with the system by their own initiative. Thus, the dialogues appeared as a need of the users for the information that the system provides. Additionally, most of the evaluation approaches cover interaction parameters and quality judgement independently. On the contrary, our results measure the relationship between these two factors, determining the statistically significant relationships. To do so, several coefficients adapted to the type of information processed have been employed, such as *Pearson correlation* coefficients, *ANOVA* studies, *tau-b*, *rho* and *Eta-square* coefficients. Thus, our study provides **new empirical evidence** which is more relevant to predict real systems behaviour, yielding very interesting results on the criteria that affected more deeply the user satisfaction and task success, the impact of users knowledge and experience or the suitability of different dialogue management initiatives. These interesting empirical relationships can be taken into account to enhance system development and also for the evaluation of the systems performance and usability.

6.2 Future work

6.2.1. Recognition of non-acted emotions

The future work includes evaluating the proposed technique using other corpora, to guarantee its independence of the application domain. Additionally, we are currently integrating the emotion recognizer into the UAH system architecture. This module will work in parallel with the speech recognizer, accepting the voice signal as an input that comes from the telephony card, and obtaining the emotion recognized, which will be sent to the dialogue manager. Then, the dialogue manager will adapt its behaviour to the emotion recognized. The initial objective is to use only the emotions described in the Thesis: *angry*, *bored* and *doubtful*. Dialogue management for doubtful users will require providing them a more detailed help and using system-directed initiatives, so that it will be clear to the user what he is required to do in the interaction step. On the other hand, when the user is bored, the optimal strategy could be to make the system prompts shorter, change the prosody of the synthesized speech, and employ more implicit confirmations in which the user will not be explicitly prompted to confirm the information he provides. For the *angry* emotional state, error recovery strategies must be considered and incorporated to the system, with some feedback about the possible on misunderstandings.

6.2.2. Cross-lingual adaptation of speech recognizers

Our immediate future objective is to compare the adapted recognizers with a plain Spanish and Slovak system built from the scratch. We have already carried out several experiments with the MS Vista Spanish recognizer. The results of this evaluation show an error rate of around 20%, as against 24% in the adapted system. This indicates that the adapted system is not too far off a native-language system. Further experiments on this topic will be carried out in the near future.

The promising findings described in Chapter 4 encourage us to consider the future application of the proposed cross-language phonetic adaptation for languages with very small speech and linguistic resources. A particularly interesting task would be to study the suitability of the proposed

cross-lingual adaptation technique for minority languages that do not belong to the Indo-European family.

6.2.3. Field evaluation of spoken dialogue systems

Besides the dialogues used in the Thesis, we have acquired another corpus under laboratory conditions, i.e. by means of recruited users employing the UAH system using predefined scenarios. Our objective is to carry out an evaluation of the results obtained with this method, and compare them with the results of the field evaluation described in Chapter 5.

We also believe that statistical analysis such as the ones presented in the Thesis lead to interesting empirical relationships that can be taken into account to enhance system development and evaluation. Such studies can serve to evaluate systems as a whole instead of individual components. Another future line of work will focus on incorporating factor analysis studies to group different criteria to obtain the major trends necessary to optimize the performance and foster user satisfaction. For this purpose a more extensive list of criteria will be compiled. Once the factors are computed, they will be analysed to obtain their interrelationships so as to be able to build a criteria taxonomy that can then be compared with other state-of-the-art taxonomies, for example the Quality-Of-Service proposed by (Möller, 2002).

New criteria will be introduced and studied to evaluate the affective intelligence. In this way, it will be possible to measure the benefits of adding the proposed emotional intelligence mechanism to the UAH dialogue system (Section 6.2.1), both objectively (e.g. in terms of task success) and subjectively (considering the users opinions).

A
Publications

The research described in the Thesis has been published in the national and international conferences and journals listed below.

There are several publications describing different stages of the design and development of the UAH spoken dialogue system (Chapter 2). The next paper presents the UAH spoken dialogue system, describing all its modules and the innovative approaches employed for its development.

- (Callejas and López-Cózar, 2005b) *Callejas, Z., López-Cózar, R., 2005. Implementing modular dialogue systems: a case study. In: Proc. of Applied Spoken Language Interaction in Distributed Environments (ASIDE 05). Aalborg, Denmark.*

The following papers correspond to a description and a demonstration of the GAG module for the UAH system. They correspond to the first version of the module in which the grammar rules were designed and generated dynamically using a web interface, but the automatic update with the database changes was still not developed:

- (Callejas and López-Cózar, 2005c) *Callejas, Z., López-Cózar, R., 2005. Nueva técnica de generación automática de gramáticas para sistemas de diálogo. Procesamiento del Lenguaje Natural (35), 205 - 212.*

- (Callejas and López-Cózar, 2005a) *Callejas, Z., López-Cózar, R., 2005. GAG: Generación automática de gramáticas en un sistema conversacional de interacción oral. Procesamiento del Lenguaje Natural (35), 457 - 458.*

The final version of the GAG tool, in which the process of grammar creation and updating was fully automatic, was presented in the following paper, in which the TGC technique was also fully designed, developed and evaluated.

- (Callejas and López-Cózar, 2007a) *Callejas, Z., López-Cózar, R., 2007. Automatic creation of ASR grammar rules for unknown vocabulary applications. In: Proc. of 8th International Workshop on Electronics, Control, Modelling, Measurement and Signals (ECMS). Liberec, Czech Republic. pp.50-55*

Regarding the research on emotion recognition described in Chapter 3, the next paper presents a state on the art on emotion recognition paying special attention to the latest international projects in the area, putting our research into context:

- (Callejas and López-Cózar, 2007c) *Callejas, Z., López-Cózar, R., 2007. Emotion recognition for spoken dialogue systems. In: Proc. of I Simposio en Desarrollo de Software (SDS 2007). Granada, Spain. pp. 59-68*

A complete study on Kappa coefficients and how they are affected by the skewness of the corpora of non-acted emotions is presented in:

- (Callejas and López-Cózar, 2008b) *Callejas, Z., López-Cózar, R., 2008. On the use of kappa coefficients to measure the reliability of the annotation of non-acted emotions. In: Proc. of 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems (PIT08). Kloster Irsee, Germany. To be published in LNCS.*

The next paper presents the first stage on the development of the proposed automatic emotion recognition approach using contextual information. It presents the preliminar version of the two-steps algorithm described in Section 3.5:

- (Callejas and López-Cózar, 2007b) *Callejas, Z., López-Cózar, R., 2007. Decisive factors in the annotation of emotions for spoken dialogue systems. Advances in Soft Computing (45), 747 - 754.*

The main results of the research described in Chapter 3 are presented in the following paper. It describes the work on both human and automatic emotion recognition using contextual information, including only the optimal version of the two-steps method for machine-learned recognition. The state of the art and kappa coefficients study are less detailed than the ones presented in the previously described paper.

- (Callejas and López-Cózar, 2008a) *Callejas, Z., López-Cózar, R., 2008. Influence of contextual information in emotion annotation for spoken dialogue systems. Speech Communication Vol. 50 (5), 416-433.*

Regarding the research on the cross-lingual adaptation of a Czech speech recognizer described in Chapter 4, the following paper corresponds to a demonstration of the Spanish version of the MyVoice system recognizing Spanish with the adapted Czech speech recognizer (Chapter 4).

- (Callejas et al., 2007) *Callejas, Z., Nouza, J., Cerva, P., López-Cózar, R., 2007. Myvoice goes spanish. Cross-lingual adaptation of a voice-controlled PC tool for handicapped people. Procesamiento del Lenguaje Natural (39), 277 - 278.*

Finally, the research done on field evaluation of SDSs (Chapter 5) is presented in.

- (Callejas and López-Cózar, 2008c) *Callejas, Z., López-Cózar, R. Relations between de-facto criteria in the evaluation of a spoken dialogue system. Speech Communication. In press, available online since 15th April 2008. DOI: 10.1016/j.specom.2008.04.004*

Bibliography

Adell, J., Bonafonte, A., Escudero, D., 2005. Analysis of prosodic features towards modelling of emotional and pragmatic attributes of speech. Procesamiento de Lenguaje Natural 35, 277–284.

Ai, H., Litman, D., Forbes-Riley, K., Rotaru, M., Tetreault, J., Purandare, A., 2006. Using systems and user performance features to improve emotion detection in spoken tutoring dialogs. In: Proc. of the 9th Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pp. 797–800.

Alexandersson, J., Becker, T., 2001. Overlay as the basic operation for discourse processing in a multimodal dialogue system. In: Proc. of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems. Seattle, USA.

Andeani, G., Fabbrizio, D. D., Gilbert, M., Gillick, D., Hakkani-Tur, D., Lemon, O., 2006. Let's DISCOH: Collecting an Annotated Open Corpus with Dialogue Acts and Reward Signals for Natural Language Helpdesks. In: Proc. of IEEE 2006 Workshop on Spoken Language Technology (SLT'06). Palm Beach, Aruba, pp. 218–221.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Proc. of the 7th International Conference on Spoken Language Processing (Interspeech'02-ICSLP). Denver, USA, pp. 2037–2040.

Artstein, R., Poesio, M., 2005. $kappa_3$ = alpha (or beta). Tech. rep., University of Essex.

AUBADE, 2005. `http://www.aubade-group.com/`.

Augmented Multiparty Interaction Project, 2007. `http://www.amiproject.org/`.

Baggia, P., Castagneri, G., Danieli, M., 2000. Field trials of the Italian ARISE train timetable system. Speech Communication 31, 355–367.

Bangalore, S., Hakkani-Tur, D., Tur, G., 2006. Introduction to the Special Issue on Spoken Language Understanding in Conversational Systems. Speech Communication 48, 233–3238.

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russel, M., Wong, M., 2004. Towards multilingual speech recognition using data driven source/target acoustical units association. In: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04). Montreal, Quebec, Canada, pp. 521–524.

Baudoin, F., Bretier, P., Corruble, V., 2005. A dialogue agent with adaptive and proactive capabilities. In: Proc. of IEEE/WIC/ACM International Conference on Intelligent Agent Technology. Compiegne, France, pp. 293–296.

Bayeh, R., Lin, S., Chollet, G., Mokbel, C., 2004. You stupid tin box! - children interacting with the aibo robot: a cross-linguistic emotional speech corpus. In: Proc. of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal, pp. 171–174.

Becker, T., Gerstenberger, C., Kruijff-Korbayova, I., Korthauer, A., Pinkal, M., Pitz, M., Poller, P., Schehl, J., 2006. Natural and intuitive multimodal dialogue for In-Car Applications: The SAMMIE System. In: Proc. of the 4th European Conference of Prestigious Applications of Intelligent Systems (PAIS'06). Riva del Garda, Italy, pp. 612–616.

Beringer, N., Kartal, U., Louka, K., Schiel, F., Tük, U., 2002. PROMISE: A Procedure for Multimodal Interactive System Evaluation. In: Proc. of the Workshop on Multimodal Resources and Multimodal Systems Evaluation. Las Palmas de Gran Canaria, Spain, pp. 77–80.

Bernsen, N., Dybkjaer, L., 1997. The DISC project. In: ELRA Newsletter. Vol. 2(2). pp. 6–8.

Bernsen, N., Dybkjaer, L., Dybkjaer, H., 1994. A dedicated task-oriented dialogue theory in support of spoken language dialogue system design. In:

Proc. of the 3rd International Conference on Spoken Language Processing (ICSLP'94). Yokohama (Japan), pp. 875–878.

Bernsen, N. O., Dybkjaer, L., 2000. A methodology for evaluating spoken language dialogue systems and their components. In: Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC'00). Athens, Greece, pp. 183–188.

Bickmore, T., Giorgino, T., 2004. Some novel aspects of health communication from a dialogue systems perspective. In: Proc. of AAAI Fall Symposium on Dialogue Systems for Health Communication. Washington DC, USA, pp. 275–291.

Bickmore, T., Giorgino, T., 2006. Health dialog systems for patients and consumers. Journal of Biomedical Informatics 39, 556–571.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Springer.

Boehner, K., DePaula, R., Dourish, P., Sengers, P., 2007. How emotion is made and measured. International Journal of Human-Computer Studies, Special Issue on Evaluating Affective Interactions 65 (4), 275–291.

Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Tech. rep., Institute of Phonetic Sciences, University of Amnsterdam.

Bohus, D., Grau, S., Huggins-Daines, D., Keri, V., Krishna, G., Kumar, R., Raux, A., Tomko, S., 2007. Conquest - an Open-Source Dialog System for Conferences. In: Proc. of Human Language Technologies'07: The Conference of the North American Chapter of the Association for Computational Linguistics. Rochester, NY, USA, pp. 9–12.

Bohus, D., Rudnicky, A., 2002. LARRI: A Language-Based Maintenance and Repair Assistant. In: Proc. of Multi-Modal Dialogue in Mobile Environments Conference (IDS'02). Kloster Irsee, Germany, pp. 203–218.

Bonaventura, P., Gallocchio, F., Micca, G., 1997. Multilingual speech recognition for flexible vocabularies. In: Proc. of 5th European Conference on Speech Communication and Technology (Eurospeech 1997). Rhodes, Greece, pp. 355–358.

Bos, J., Klein, E., Lemon, O., Oka, T., 1999. The verbmobil prototype system - a software engineering perspective. Journal of Natural Language Engineering 5(1), 95–112.

Boves, L., Os, E. D., 2002. Multimodal services, a MUST for UMTS. Tech. rep., EURESCOM.

Brown, M. K., 1999. Grammar Representation Requirements for Voice Markup Languages. Tech. rep., W3C.

Burkhardt, F., van Ballegooy, M., Englert, R., Huber, R., 2005. An emotion-aware voice portal. In: Proc. of Electronic Speech Signal Processing. Prague, Czech Republic, pp. 123–131.

Cahill, L., Tiberius, C., 2002. Cross-linguistic phoneme correspondences. In: Proc. of 19th International Conference on Computational Linguistics (COLING'02). Taipei, Taiwan, pp. 1–5.

Callas - Conveying Affectiveness in Leading-Edge Living Adaptive Systems, 2007. http://www.callas-newmedia.eu/.

Callejas, Z., López-Cózar, R., 2005a. GAG: Generación automática de gramáticas en un sistema conversacional de interacción oral. Procesamiento del Lenguaje Natural 35, 457–458.

Callejas, Z., López-Cózar, R., 2005b. Implementing modular dialogue systems: a case study. In: Proc. of Applied Spoken Language Interaction in Distributed Environments (ASIDE'05). Aalborg, Denmark.

Callejas, Z., López-Cózar, R., 2005c. Nueva técnica de generación automática de gramáticas para sistemas de diálogo. Procesamiento del Lenguaje Natural 35, 205–212.

Callejas, Z., López-Cózar, R., 2007a. Automatic creation of ASR grammar rules for unknown vocabulary applications. In: Proc. of the 8th International workshop on Electronics, Control, Modelling, Measurement and Signals (ECMS'07). Liberec, Czech Republic, pp. 50–55.

Callejas, Z., López-Cózar, R., 2007b. Decisive factors in the annotation of emotions for spoken dialogue systems. Advances in Soft Computing 45, 747–754.

Callejas, Z., López-Cózar, R., 2007c. Emotion recognition for spoken dialogue systems. In: Proc. of I Simposio en Desarrollo de Software (SDS'07). Granada, Spain, pp. 59–68.

Callejas, Z., López-Cózar, R., 2008a. Influence of contextual information in emotion annotation for spoken dialogue systems. Speech Communication 50 (5), 416–433.

Callejas, Z., López-Cózar, R., 2008b. On the use of kappa coefficients to measure the reliability of the annotation of non-acted emotions. In: Proc. of 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems (PIT'08). Kloster Irsee, Germany, to be published in LNCS.

Callejas, Z., López-Cózar, R., 2008c. Relations between de-facto criteria in the evaluation of a spoken dialogue system. Speech Communication In Press. Available online since 15th April 2008. DOI: 10.1016/j.specom.2008.04.004.

Callejas, Z., Nouza, J., Cerva, P., López-Cózar, R., 2007. Myvoice goes spanish. cross-lingual adaptation of a voice-controlled pc tool for handicapped people. Procesamiento del Lenguaje Natural 39, 277–278.

Camurri, A., Mazzarino, B., Volpe, G., 2004. Expressive interfaces. Cognition, Technology and Work 6 (1), 15–22.

Carreiras, M., García-Albea, J. E., Sebastián-Gallés, N., 1996. Language processing in Spanish. Lawrence Erlbaum Associates.

Castagneri, G., Baggia, P., Danieli, M., 1998. Field trials of the Italian ARISE train timetable system. In: Proc. of the Interactive Voice Technology for Telecommunications Applications Workshop (IVTTA'98). pp. 97–102.

Catizone, R., Setzer, A., Wilks, Y., 2003. Multimodal Dialogue Management in the COMIC Project. In: Proc. of EACL'03 Workshop on Dialogue Systems: interaction, adaptation, and styles of management. Budapest, Hungary, pp. 25–34.

Cerrato, L., 2002. A comparison between feedback strategies in human-to-human and human-machine communication. In: Proc. of the 8th International Conference on Spoken Language Processing (ICSLP 2002). Vol. 2. Denver, USA, pp. 557–560.

Cerva, P., Nouza, J., 2007. Design and development of voice controlled aids for motor-handicapped persons. In: Proc. of the 11th International Conference on Spoken Language Processing (Interspeech'07-Eurospeech). Antwerp, Belgium, pp. 2521–2524.

Chambers, N., Allen, J., 2004. Stochastic language generation in a dialogue system: Toward a domain independent generator. In: Proc. of the 5th SIGdial Workshop on Discourse and Dialogue. Boston, USA, pp. 9–18.

Cheyer, A., Julia, L., 1995. Multimodal maps: An agent based approach. In: Proc. of International Conference on Cooperative Multimodal Cooperation. Eindhoven, Holland, pp. 111–121.

CHIL - Computers In the Human Interaction Loop, 2007. `http://chil.server.de/servlet/is/101/`.

Chu, S.-W., O'Neill, I., Hanna, P., McTear, M., 2005. An approach to multi-strategy dialogue management. In: Proc. of 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech). Lisbon, Portugal, pp. 865–868.

Cicchetti, D. V., Feinstein, A. R., 1990. High agreement but low Kappa: II. Resolving the paradoxes. Journal of Clinical Epidemiology 43 (6), 551–558.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20 (3), 37–46.

Cohen, J., 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70 (4), 213–220.

Cole, R., Hirschman, L., Atlas, L., Beckman, M., Biermann, A., Bush, M., Clements, M., Cohen, J., Garcia, O., Hanson, B., Hermansky, H., Levinson, S., McKeown, K., Morgan, N., Novick, D. G., Ostendorf, M., Oviatt,

S., Price, P., Silverman, H., Spitz, J., Waibel, A., Weinstein, C., Zahorian, S., Zue, V., 1995. The Challenge of Spoken Language Systems: Research Direction for the Nineties. IEEE Transactions on Speech and Audio Processing 3, 1–20.

Cole, R., Mariani, J., Uszkoreit, H., Varile, G. B., Zaenen, A., Zampolli, A., Zue, V. (Eds.), 1997. Survey of the State of the Art in Human Language Technology. Cambridge University Press.

Companions, 2007. http://nlp.shef.ac.uk/companions/.

Corradini, A., Bernsen, N., Fredriksson, M., Johanneson, L., Königsmann, J., Mehta, M., 2004. Towards believable behavior generation for embodied conversational agents. In: Proc. of the Workshop on Interactive Visualisation and Interaction Technologies (IV&IT 2004). Krakow, Poland, pp. 946–953.

Corradini, A., Mehta, M., Bernsen, N. O., Charfuelán, M., 2005. Animating an interactive conversational character for an educational game system. In: Proc. of the 2005 International Conference on Intelligent User Interfaces. San Diego, CA, USA, pp. 183–190.

CoSy Home, 2007. http://www.cs.bham.ac.uk/research/projects/cosy/.

Cowie, R., 2000. Describing the emotional states expressed in speech. In: Proc. of the International Speech Communication Association (ISCA) Workshop on Speech and Emotion. Newcastle, Northern Ireland, UK, pp. 11–18.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M., 2000. FEELTRACE: An Instrument for Recording Perceived Emotion in Real Time. In: Proc. of the International Speech Communication Association (ISCA) Workshop on Speech and Emotion. Newcastle, Northern Ireland, UK, pp. 19–24.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., 2001. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18 (1), 32–80.

Cowie, R., Schröder, M., 2005. Piecing Together the Emotion Jigsaw. Lecture Notes on Computer Science 3361/2005, 305–317.

Craggs, R., Wood, M. M., 2003. Annotating emotion in dialogue. In: Proc. of the 4th SIGdial Workshop on Discourse and Dialogue. Sapporo, Japan, pp. 218–225.

Critchley, H. D., Rotshtein, P., Nagai, Y., ODoherty, J., Mathias, C. J., Dolana, R. J., 2005. Activity in the human brain predicting differential heart rate responses to emotional facial expressions. NeuroImage 24, 751–762.

Cuayáhuitl, H., Renals, S., Lemon, O., Shimodaira, H., 2006. Reinforcement learning of dialogue strategies with hierarchical abstract machines. In: Proc. of IEEE/ACL Workshop on Spoken Language Technology (SLT'06). Palm Beach, Aruba, pp. 182–186.

Dale, R., September 2003. Next-generation spoken dialog systems. Technology Trends Seminar Series, Macquarie University, http://www.ict.csiro.au/MU/Trends/2003.htm.

DARPA, 1992. Speech and Natural Language Workshop. Defense Advanced Research Projects Agency (DARPA), San Mateo, USA.

DARPA, 1994. Speech and Natural Language Workshop. Defense Advanced Research Projects Agency (DARPA), San Mateo, USA.

Davies, M., Fleiss, J. L., 1982. Measuring agreement for multinomial data. Biometrics 38 (4), 1047–1051.

de Melo, C., Paiva, A., 2005. Environment expression: Expressing emotions through cameras, lights and music. In: Proc. of the 1st International Conference on Affective Computing and Intelligent Interaction (ACII'05). Beijing, China, pp. 715–722.

Degerstedt, L., Jönsson, A., 2006. LinTest, A development tool for testing dialogue systems. In: Proc. of the 9th Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pp. 489–492.

den Os, E., Boves, L., Lamel, L., Baggia, P., 1999. Overview of the ARISE project. In: Proc. of the European Conference on Speech Technology (Eurospeech'99). Budapest, Hungary, pp. 1527–1530.

Devillers, L., Maynard, H., Rosset, S., 2004. The French Media/Evalda project: the evaluation of the understanding capability of spoken language dialog systems. In: Proc. of the 4th International Conference on Language Resources and Evaluation (LREC 2004). Vol. 6. Lisbon, Portugal, pp. 2131–2134.

Devillers, L., Vidrascu, L., Lamel, L., 2005. Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18, 407–422.

DISC, 1999. DISC Final Report covering the period from 1.6.98 to 28.2.99. Deliverable D5.2. Tech. rep., The DISC Consortium.

Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P., 2003. Emotional speech: towards a new generation of databases. Speech Communication 40, 33–60.

Dunn, G., 1989. Design and analysis of reliability studies: the statistical evaluation of measurement errors. Edward Arnold.

Dybkjaer, L., Bernsen, N. O., 2000. Usability issues in spoken language dialogue systems. Natural Language Engineering 6, 243–271.

Dybkjaer, L., Bernsen, N. O., Minker, W., 2004. Evaluation and usability of multimodal spoken language dialogue systems. Speech Communication 43, 33–54.

EAGLES, 1996. Evaluation of Natural Language Processing Systems. Final report. Document EAG-EWG-PR2. Tech. rep., Center for Sprogetknologi, Copenhague, Denmark.

EUROPA - CORDIS: Community Research and Development Information Service, 2006. `http://cordis.europa.eu/`.

Farzanfar, R., Frishkopf, S., Migneault, J., Friedman, R., 2004. Telephone-linked care for physical activity: A qualitative evaluation of the use

patterns of an information technology program for patients. Journal of Biomedical Informatics 38 (3), 220–228.

Feinstein, A. R., Cicchetti, D. V., 1990. High agreement but low Kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology 43 (6), 543–549.

Fleiss, J. L., 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin 76 (5), 378–382.

Fleiss, J. L., Cohen, J., 1973. The equivalence of weighted kappa and the interclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 33, 613–619.

Fluhr, C., Schmit, D., Andrieux, C., Ortet, P., Bisson, F., Combet, V., 1999. Crosslingual interrogation of multilingual catalogs. Lecture Notes on Computer Science 1696, 294–310.

Forbes-Riley, K., Litman, D. J., 2004a. Predicting emotion in spoken dialogue from multiple knowledge sources. In: Proc. of the Human Language Technology Conference - North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL'04). Boston, USA, pp. 201–208.

Forbes-Riley, K. M., Litman, D., 2004b. Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters. In: Proc. of the Human Language Technology Conference - North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL'06). New York, USA, pp. 264–271.

Gales, M. J. F., Woodland, P. C., 1996. Mean and Variance Adaptation Within the MLLR Framework. Computer Speech and Language 10, 249–264.

Gao, Y., Gu, L., Jeff, H.-K., 2005. Portability challenges in developing interactive dialogue systems. In: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05). Vol. 5. Philadelphia, PA, USA, pp. 18–23.

Gauvain, J., 1999. The limsi sdr system for trec. In: Proc. of the 8th Text Retrieval Conference (TREC 8). Gaithersburg, Maryland, USA, pp. 475–482.

Gauvain, J. L., Lee, C. H., 1994. Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. IEEE Transactions on Speech and Audio Processing 2, 291–298.

Gebhard, P., Klesen, M., Rist, T., 2004. Coloring multi-character conversations through the expression of emotions. In: Proc. of Tutorial and Research Workshop on Affective Dialogue Systems. Kloster Irsee, Germany, pp. 128–141.

Gerfen, C., 2002. Andalusian codas. Probus 14, 247–277.

Geutner, P., Steffens, F., Manstetten, D., 2002. Design of the VICO Spoken Dialogue System: Evaluation of User Expectations by Wizard-of-Oz experiments. In: Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC'02). Las Palmas de Gran Canaria, Spain, pp. 385–400.

Gibbon, D., Mertins, I., Moore, R., 2000. Handbook of Multimodal and Spoken Dialogue Systems. Resources, Terminology and Product Evaluation. Kluwer Academic Publishers (Kluwer International Series in Engineering and Computer Science, 565).

Gibbon, D., Moore, R., Winski, R., 1997. Handbook of Standards and Resources for Spoken Language Systems. Walter de Gruyter.

Glass, J., Flammia, G., Goodine, D., Phillips, M., Polifroni, J., Sakai, S., Seneff, S., Zue, V., 1995. Multilingual spoken-language understanding in the MIT Voyager system. In: Speech Communication. Vol. 17. pp. 1–18.

Glass, J. R., 1999. Challenges for spoken dialogue systems. In: Proc. of the Automatic Speech Recognition and Understanding Workshop (ASRU'99). Keystone, Colorado, USA.

González, G. M., 1999. Bilingual computer-assisted psychological assessment: An innovative approach for screening depression in chicanos/latinos. Tech. Rep. 39, University of Michigan, Ann Arbor, USA.

Griol, D., 2007. Desarrollo y Evaluación de diferentes Metodologías para la Gestión Automática del Diálogo. Ph.D. thesis, Universidad Politécnica de Valencia, Valencia, Spain.

Gupta, N., Tur, G., Hakkani-Tur, D., Bangalore, S., Riccardi, G., Gilbert, M., 2006. The AT&T spoken language understanding system. IEEE Transactions on Audio, Speech and Language processing 14, 213–222.

Gustafson, J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D., Wirén, M., 2000. AdApt - a multimodal conversational dialogue system in an apartment domain. In: Proc. of the 6th International Conference on Spoken Language Processing (ICSLP'00). Vol. 2. Beijing, China, pp. 134–137.

Gut, U., Bayerl, P. S., 2004. Measuring the reliability of manual annotations of speech corpora. In: Proc. of the 2nd International Conference on Speech Prosody (SP'04). Nara, Japan, pp. 565–568.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182.

Haas, J., Gallwitz, F., Horndasch, A., Huber, R., Warnke, V., 2005. Telephone-based speech dialog systems. Lecture Notes on Computer Science 3663, 125–132.

Hall, L., Woods, S., Aylett, R., Paiva, A., Newall, L., 2005. Achieving empathic engagement through affective interaction with synthetic characters. In: Proc. of the 1st International Conference on Affective Computing and Intelligent Interaction (ACII'05). Beijing, China, pp. 731–738.

Hamerich, S., de Córdoba, R., Schless, V., d'Haro, L., Schubert, V., Kocsis, O., Igel, S., Pardo, J. M., 2004. The GEMINI Platform: Semi-Automatic Generation of Dialogue Applications. In: Proc. of the 8th International Conference on Spoken Language Processing (Interspeech-ICSLP). Vol. 4. Jeju, Corea, pp. 2629–2632.

Hanna, P., O'Neill, I., Wootton, C., McTear, M., 2007. Promoting extension and reuse in a spoken dialog manager: an evaluation of the Queen's Communicator. ACM Transactions on Speech and Language Processing 4.

Hansen, J. H. L., 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Speech communication 20 (2), 151–170.

Hardy, H., Biermann, A., Inouye, R., McKenzie, A., Strzalkowski, T., Ursu, C., Webb, N., Wu, M., 2006. The Amitiés system: Data-driven techniques for automated dialogue. Speech Communication 48, 354–373.

Hartikainen, M., Salonen, E.-P., Turunen, M., 2004. Subjective Evaluation of Spoken Dialogue Systems Using SERVQUAL Method. In: Proc. of 8th International Conference on Spoken Language Processing (Interspeech'04-ICSLP). Jeju Island, Korea, pp. 2273–2276.

Haseel, L., Hagen, E., 2005. Adaptation of an automotive dialogue system to users' expertise. In: Proc. of 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech). Lisbon, Portugal, pp. 222–226.

Heim, J., Nilsson, E. G., Skjetne, J. H., 2007. User Profiles for Adapting Speech Support in the Opera Web Browser to Disabled Users. Lecture Notes on Computer Science 4397, 154–172.

Higashinaka, R., Miyazaki, N., Nakano, M., Aikawa, K., 2004. Evaluating discourse understanding in spoken dialogue systems. ACM Transactions on Speech and Language Processing 1, 1–20.

Holmes, W., Huckvale, M., 1994. Why have HMMs been so successful for automatic speech recognition and how might they be improved? In: Speech Hearing and Language. pp. 1875–1878.

Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., Nogueiras, A., 2002. Interface databases: Design and collection of a multilingual emotional speech database. In: Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC'02). Las Palmas de Gran Canaria, Spain, pp. 385–400.

Hu, Y., Loizou, P. C., 2007. Subjective comparison and evaluation of speech enhacement algorithms. Speech Communication 49, 588–601.

Hubal, R. C., Frank, G. A., Guinn, C. I., 2000. AVATALK Virtual Humans for Training with Computer Generated Forces. In: Proc. of the 9th Conference on Computer Generated Forces and Behavioral Representation. Orlando, Florida, USA, pp. 617–623.

Humaine emotion-research.net, 2007. http://emotion-research.net/.

Hurtig, T., 2004. Visualization and multimodality: a mobile multimodal dialogue system for public transportation navigation evaluated. In: Proc. of the 8th Conference on Human-computer interaction with mobile devices and services (MobileHCI'06). Helsinki, Finland, pp. 251–254.

Interspeech 2007 Special Session on Speech and language technology for less-resourced languages, 2007. http://www.interspeech2007.org/Technical/less_resourced_languages.php.

Intrepid Project (A virtual reality intelligent multi-sensor wearable system for phobias' treatment), 2004. http://www.intrepid-project.org/.

Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J. M., Bernadas, D., Oliver, J. M., Tena, D., Longhi, L., 2000. Validation of an acoustical modelling of emotional expression in spanish using speech synthesis techniques. In: Proc. of the International Speech Communication Association (ISCA) Workshop on Speech and Emotion. Newcastle, Northern Ireland, UK, pp. 161–166.

J. Gustafson, N. L., Lundeberg, M., 1999. The August spoken dialogue system. In: Proc. of the 6th European Conference on Speech Communication and Technology (EuroSpeech'99). Budapest, Hungary, pp. 1151–1154.

Johnstone, T., 1996. Emotional speech elicited using computer games. In: Proc. of the 4th International Conference on Spoken Language Processing (ICSLP 1996). Vol. 3. Philadelphia, PA, pp. 1985–1988.

Jokinen, K., 2003. Natural interaction in spoken dialogue systems. In: Proc. of the Workshop Ontologies and Multilinguality in User Interfaces. Crete, Greece, pp. 730–734.

Jokinen, K., Kanto, K., Rissanen, J., 2004. Adaptative User Modelling in AthosMail. Lecture Notes on Computer Science 3196, 149–158.

Jurafsky, D., Martin, J. H., 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall.

Kacic, Z., 1999. Advances in spoken dialogue systems development. In: Proc. of IEEE International Symposium on Industrial Electronics (ISIE'99). Vol. 1. Bled, Slovenia, pp. 169–172.

Kirchhoff, K., Bilmes, J. A., 1999. Statistical acoustic indications of coarticulation. In: Proc. of International Congress of Phonetic Sciences. San Francisco, California, USA, pp. 1729–1732.

Kirchhoff, K., Vergyri, D., 2005. Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition. Speech Communication 46, 37–51.

Krippendorff, K., 2003. Content Analysis: An Introduction to its Methodology. Sage Publications, Inc.

Krsmanovic, F., Spencer, C., Jurafsky, D., Ng, A. Y., 2006. Have we meet? MDP Based Speaker ID for Robot Dialogue. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pp. 461–464.

Kumar, C. S., Mohandas, V. P., Haizhou, L., 2005. Multilingual speech recognition: A unified approach. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech). Lisbon, Portugal, pp. 3357–3360.

Kwon, O., Yoo, K., Suh, E., 2005. ubiES: An Intelligent Expert System for Proactive Services Deploying Ubiquitous Computing Technologies. In: Proc. of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05). Hawaii, pp. 85–86.

Lamel, L., Bennacef, S., Gauvain, J., Dartigues, H., Temem, J., 2002. User evaluation of the MASK kiosk. Speech Communication 38 (1-2), 131–139.

Lamel, L., Minker, W., Paroubek, P., 2000a. Towards best practice in the development and evaluation of speech recognition components of a spoken language dialog system. Natural Language Engineering 6 (3-4), 305–322.

Lamel, L., Rosset, S., Gauvain, J., Bennacef, S., Garnier-Rizet, M., Prouts, B., 2000b. The LIMSI ARISE system. Speech Communication 31, 339–353.

Landis, J. R., Koch, G. G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Langner, B., Black, A., 2005. Using speech in noise to improve understandability for elderly listeners. In: Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'05). San Juan, Puerto Rico, pp. 392–396.

Lantz, C. A., Nebenzahl, E., 1996. Behavior and interpretation of the κ statistic: Resolution of the two paradoxes. Journal of Clinical Epidemiology 49 (4), 431–434.

Larsen, L. B., 2003. Issues in the evaluation of spoken dialogue systems using objective and subjective measures. In: Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03). St. Thomas, U.S. Virgin Islands, USA, pp. 209–214.

Lee, C., Yoo, S. K., Park, Y. J., Kim, N. H., Jeong, K. S., Lee, B. C., 2005. Using Neural Network to Recognize Human Emotions from Heart Rate Variability and Skin Resistance. In: Proc. of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'05). Shanghai, China, pp. 5523–5525.

Lee, C. M., Narayanan, S. S., 2005. Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing 13 (2), 293–303.

Legris, P., Ingham, J., Collerette, P., 2003. Why do people use information technology: A critical review of the technology acceptance model. Information and Management 40, 191–204.

Lehmann, S., Oepen, S., Regnier-Prost, S., Netter, K., Lux, V., Klein, J., Falkedal, K., Fouvry, F., Estival, D., Dauphin, E., Compagnion, H., Baur, J., Balkan, L., Arnold, D., 1996. TSNLP — Test Suites for Natural Language Processing. In: Proc. of the 16th International Conference on Computational Linguistics (COLING'96). Copenhagen, Denmark, pp. 711–716.

Lemon, O., Bracy, A., Gruenstein, A., Peters, S., 2001. The Witas Multi-Modal Dialogue System. In: Proc. of Eurospeech 2001. Aalborg, Denmark, pp. 1559–1562.

Lemon, O., Georgila, K., Henderson, J., 2006. Evaluating Effectiveness and Portability of Reinforcement Learned Dialogue Strategies with real users: the TALK TownInfo Evaluation. In: Proc. of IEEE-ACL Workshop on Spoken Language Technology (SLT 2006). Palm Beach, Aruba, pp. 178–181.

Ligozat, A.-L., Grau, B., Robba, I., Vilnat, A., 2006. Evaluation and improvement of cross-lingual question answering strategies. In: Proc. of the 11th EACL Workshop on Multilingual Question Answering (MLQA'06). Trento, Italy, pp. 23–30.

Lines, L., Hone, K. S., 2002. Older adults' evaluations of speech output. In: Proc. of the 5th International ACM SIGCAPH Conference on Assistive Technologies (ASSETS'02). Edinburgh, Scotland, pp. 170–177.

Liscombe, J., Riccardi, G., Hakkani-Tür, D., 2005. Using context to improve emotion detection in spoken dialog systems. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech). Lisbon, Portugal, pp. 1845–1848.

Litman, D. J., Forbes-Riley, K., 2006. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. Speech communication 48 (5), 559–590.

Litman, D. J., Pan, S., 2002. Designing and evaluating an adaptive spoken dialogue system. User modelling and user-adapted interaction 12, 111–137.

Litman, D. J., Silliman, S., 2004. ITSPOKE: An Intelligent Tutoring Spoken Dialogue System. In: Proc. of the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL). Boston, USA, pp. 233–236.

López-Cózar, R., Araki, M., 2005. Spoken, Multilingual and Multimodal Dialogue Systems. Development and Assessment. John Wiley and Sons.

LREC 2008 Workshop on Collaboration: interoperability between people in the creation of language resources for less-resourced languages, 2008. http://ixa2.si.ehu.es/saltmil/en/home/.

MagiCster Project Pages, 2007. http://www.ltg.ed.ac.uk/magicster/.

Mahlke, S., 2006. Emotions and EMG measures of facial muscles in interactive contexts. In: Proc. of the 2006 Conference on Human Factors in Computer Systems (CHI'06). Montreal, Canada.

Mangold, H., 2001. Speech technology in reality - Applications, their challenges and solutions. In: Proc. of the 4th International Conference on Text, Speech and Dialogue (TSD'01). Pilsen, Czech Republic, pp. 197–200.

Martinovski, B., Traum, D., 2003. Breakdown in human-machine interaction: the error is the clue. In: Proc. of the ISCA tutorial and research workshop on Error handling in dialogue systems. Chateau d'Oex, Vaud, Switzerland, pp. 11–16.

Martín-Valdivia, M. T., Martínez-Santiago, F., Ureña-López, L., 2005. Merging strategy for cross-lingual information retrieval systems based on learning vector quantization. Neural Processing Letters 22 (2), 149–161.

Mattasoni, M., Omologo, M., Santarelli, A., Svaizer, P., 2002. On the Joint Use of Noise Reduction and MLLR Adaptation for In-Car Hands-Free Speech Recognition. In: Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02). Orlando, USA, pp. 289–292.

McTear, M. F., 2004. Spoken dialogue technology. Springer.

MEGA, 2001. http://www.megaproject.org/.

Melin, H., Sandell, A., Ihse, M., 2001. Ctt-bank: A speech controlled telephone banking system - an initial evaluation. In: TMH-QPSR. Vol. 1. pp. 1–27.

Menezes, P., Lerasle, F., Dias, J., Germa, T., 2007. Towards an interactive humanoid companion with visual tracking modalities. International Journal of Advanced Robotic Systems, 48–78.

Mihalcea, R., Leong, B., 2006. Toward communicating simple sentences using pictorial representations. In: Proc. of the 7th Conference of the Association for Machine Translation in the Americas (AMTA'06). Boston, MA, USA, pp. 119–127.

Minker, W., 1998. Stochastic versus rule-based speech understanding for information retrieval. Speech Communication 25(4), 223–247.

Minker, W., Albalate, A., Bühler, D., Pittermann, A., Pittermann, J., Strauss, P.-M., Zaykovskiy, D., 2006a. Recent trends in spoken language dialogue systems. In: ITI 4th International Conference on Information and Communications (ICICT'06). Montreal, Canada, pp. 1–16, invited paper.

Minker, W., Haiber, U., Heisterkamp, P., Scheible, S., 2004a. The Seneca Spoken Language Dialogue System. In: Speech Communication. Vol. 43. pp. 1–2.

Minker, W., Haiber, U., Heisterkamp, P., Scheible, S., 2004b. The SENECA spoken language dialogue system. Speech Communication 43, 89–102.

Minker, W., Pittermann, J., Pittermann, A., Strauss, P.-M., Bühler, D., 2006b. Next-generation human-computer interfaces - Towards intelligent, adaptive and proactive spoken language dialogue systems. In: Proc. of the 2nd IET International Conference on Intelligent Environments (IE'06). Vol. 1. Athens, Greece, pp. 213–219.

Mäkelä, K., Salonen, E., M., T., Hakulinen, J., Raisamo, R., 2001. Evaluating the User Interface of a Ubiquitous Computing system Doorman. In: Proc. of the 3rd International Conference on Ubiquitous Computing (Ubicomp'01). Atlanta, USA.

Möller, S., 2002. A new taxonomy for the quality of telephone services based on spoken dialogue systems. In: Proc. of the 3rd Workshop on Discourse and Dialogue (SIGDial'02). Philadelphia, USA, pp. 142–153.

Möller, S., 2005. Quality of telephone-based spoken dialogue systems. Springer.

Möller, S., Smeele, P., Boland, H., Krebber, J., 2007. Evaluating spoken dialogue systems according to de-facto standards: A case study. Computer Speech and Language 21, 26–53.

Montero, J. M., Gutiérrez-Arriola, J., Enríquez, E., Pardo, J. M., 1999. Analysis and modelling of emotional speech in Spanish. In: Proc. of the 14th International Conference of Phonetic. San Francisco, USA, pp. 957–960.

Montoro, G., Alamán, X., Haya, P. A., 2004. Spoken interaction in intelligent environments: a working system. In: Ferscha, A., Hoertner, H., Kotsis, G. (Eds.), Advances in Pervasive Computing. Austrian Computer Society (OCG), pp. 747–754.

Montoro, G., Haya, P. A., Alamán, X., López-Cózar, R., Callejas, Z., 2006. A proposal for an XML definition of a dynamic spoken interface for ambient intelligence. In: International Conference on Intelligent Computing (ICIC'06). Kunming, China, pp. 711–716.

Morgan, N., Fosler, E., Mirghafori, N., 1997. Speech recognition using on-line estimation of speaking rate. In: Proc. of the 5th European Conference on Speech Communication and Technology (Eurospeech'97). Rhodes, Greece, pp. 2079–2082.

Morrison, D., Wang, R., Silva, L. C. D., 2007. Ensemble methods for spoken emotion recognition in call-centers. Speech communication 49, 98–112.

Mostow, J., 2008. Experience from a reading tutor that listens: Evaluation purposes, excuses, and methods. In: Kinzer, C., Verhoeven, L. (Eds.), Interactive Literacy Education: Facilitating Literacy Environments Through Technology. New York: Lawrence Erlbaum Associates, Taylor and Francis, pp. 117–148.

Mourão, M., Cassaca, R., Mamede, N., 2004. An Independent Domain Dialogue System Through a Service Manager. In: Proc. of the 4th International Confernce on Advances in Natural Language Processing. Alicante, Spain, pp. 161–171.

Nakamura, S., Markov, K., Jitsuhiro, T., Zhang, J.-S., Yamamoto, H., Kikui, G., 2004. Multi-lingual speech recognition system for speech-to-speech translation. In: Proc. of 8th International Conference on Spoken Language Processing (Interspeech'04-ICSLP). Jeju Island, Korea, pp. 146–154.

Narayanan, S., Fabbrizio, G. D., Kamm, C., Hubbell, J., Buntschuh, B., Ruscitti, P., Wright, J., 2000. Effects of dialog initiative and multi-modal presentation strategies on large directory information access. In: Proc. of the 6th International Conference on Spoken Language Processing (ICSLP 2000). Vol. 2. Beijing, China, pp. 636–639.

NECA Project, 2005. http://www.ofai.at/research/nlu/NECA.

Nguyen, A., Wobcke, W., 2006. Extensibility and Reuse in an Agent-Based Dialogue Model. In: Proc. of IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. pp. 367–371.

NICE project - Main page, 2007. http://www.niceproject.com/.

Nielsen, J., 1994. Usability Engineering. Morgan Kaufmann Ed., San Francisco, USA.

Nieuwoudt, C., Botha, E. C., 1999. Cross-language adaptation of acoustic models in automatic speech recognition. In: Proc. of 5th Africon Conference in Africa (IEEE Africon 1999). Cape Town, South Africa, pp. 181–184.

Nieuwoudt, C., Botha, E. C., 2002. Cross-language use of acoustic information for automatic speech recognition. Speech Communication 38, 101–113.

Németh, G., Zainkó, C., 2003. Multilingual statistical text analysis, Zipf's law and Hungarian speech generation. Acta Linguistica Hungarica 49, 385–405.

Nouza, J., Nouza, T., Cerva, P., 2005. A multi-functional voice-control aid for disabled persons. In: Proc. of International Conference on Speech and Computer (SPECOM'05). Patras, Greece, pp. 715–718.

Nouza, J., Psutka, J., Uhlír, J., 1997. Phonetic Alphabet for Speech Recognition of Czech. Radioengineering 6 (4), 16–20.

Oh, A., Rudnicky, A., 2000. Stochastic language generation for spoken dialogue systems. In: Proc. of ANLP/NAACL 2000 Workshop on Conversational Systems. Seattle, USA, pp. 27–32.

O'Neill, P., 2005. Utterance final /s/ in Andalusian Spanish. The phonetic neutralization of a phonological contrast. Language Design 7, 151–166.

Ortony, A., G. L. Clore, A. C., 1988. The cognitive structure of emotions. Cambridge University Press.

Ostendorf, M., Digalakis, V., Kimball, O., Sep 1996. From hmm's to segment models: a unified view of stochastic modeling for speech recognition. Speech and Audio Processing, IEEE Transactions on 4 (5), 360–378.

Oviatt, S., DeAngeli, A., Kuhn, K., 1997. Integration and synchronization of input modes during multimodal human-computer interaction. In: Proc. of the SIGCHI conference on Human factors in computing systems. Atlanta, Georgia, USA, pp. 415–422.

Paek, T., 2001. Empirical methods for evaluating dialog systems. In: Proc. of the Workshop on Evaluation for Language and Dialogue Systems. Vol. 9. Toulouse, France, pp. 1–8.

Park, W., Han, S. H., Park, Y. S., Park, J., Yang, H., 2007. A framework for evaluating the usability of spoken language dialogue systems (SLDSs). Lecture Notes on Computer Science 4559, 398–404.

Pérez, G., Amores, G., Manchón, P., 2006. A multimodal architecture for home control by disabled users. In: Proc. of IEEE/ACL Workshop on Spoken Language Technology (SLT). Palm Beach, Aruba, pp. 134–137.

Persia, L. D., Yanagida, M., Rufiner, H. L., Milone, D., 2007. Objective quality evaluation in blind source separation for speech recognition in a real room. Signal Processing 87, 1951–1963.

PF-STAR home page, 2004. `http://pfstar.itc.it/`.

Pfau, T., Ruske, G., 1998. Estimating the speaking rate by vowel detection. In: Proc. of IEEE International Conference on Acoustics, Speech, and

Signal Processing (ICASSP'98). Vol. 2. Seattle, Washington, USA, pp. 945–948.

Picard, R. W., 1997. Affective Computing. The MIT Press, Cambridge, Massachusetts.

Picard, R. W., Daily, S. B., 2005. Evaluating affective interactions: Alternatives to asking what users feel. In: Proc. of the 2005 Conference on Human Factors in Computer Systems (CHI'05), Workshop: Evaluating Affective Interfaces-Innovative Approaches. Portland, Oregon, USA.

Pieraccini, R., Levin, E., Eckert, W., 1997. AMICA: The AT&T mixed initiative conversational architecture. In: Proc. of European Conference on Speech Communications and Technology (Eurospeech'97). Rhodes, Greece, pp. 1875–1878.

Pitterman, J., Pitterman, A., 2006. Integrating emotion recognition into an adaptive spoken language dialogue system. In: Proc. of the 2nd IEEE International Conference on Intelligent Environments. Athens, Greece, pp. 197–202.

Plutchik, R., 1980. EMOTION: A psychoevolutionary synthesis. Harper and Row publishers.

Polzin, T., Waibel, A., 2000. Emotion-sensitive human-computer interfaces. In: Proc. of ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion. Newcastle, Northern Ireland, UK, pp. 201–206.

Prendinger, H., Mayer, S., Mori, J., Ishizuka, M., 2003. Persona effect revisited. using bio-signals to measure and reflect the impact of character-based interfaces. In: Proc. of the 4th International Working Conference on Intelligent Virtual Agents (IVA'03). Kloster Irsee, Germany, pp. 283–291.

Radev, D., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Celebi, A., Qi, H., Drabek, E., Liu, D., 2001. Evaluation of text summarization in a cross-lingual information retrieval framework. Tech. rep., Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA.

Rajman, M., Bui, T. H., Rajman, A., Seydoux, F., Trutnev, A., Quarteroni, S., 2004. Assessing the usability of a dialogue management system designed

in the framework of a rapid dialogue prototyping methodology. Acta Acustica united with Acustica 90, 1906–1111.

Raux, A., Bohus, D., Black, A. W., Eskenazi, M., 2006. Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go! Experience. In: Proc. of the 9th Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pp. 65–68.

Raux, A., Langner, B., Black, A., Eskenazi, M., 2005. Let's Go Public! Taking a Spoken Dialog System to the Real World. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech). Lisbon, Portugal, pp. 885–888.

Raux, A., Langner, B., Black, A. W., Eskenazi, M., 2003. LET'S GO: Improving Spoken Dialog Systems for the Elderly and Non-natives. In: Proc. of the European Conference on Speech Technology (Eurospeech'03). Geneva, Switzerland, pp. 753–756.

Rayner, M., Hockey, B., Renders, J.-M., Chatzichrisafis, N., Farrell, K., 2005. A voice enabled procedure browser for the International Space Station. In: 43th Annual Meeting of the Association for Computational Linguistics. Ann Arbor, USA, pp. 29–32.

Riccardi, G., Hakkani-Tür, D., 2005. Grounding emotions in human-machine conversational systems. Lecture Notes in Computer Science, 144–154.

Rich, C., Sidner, C., 1998. COLLAGEN: A Collaboration Manager for Software Interface Agents. User Modeling and User-Adapted Interaction 8, 315–350.

Robinson, S. M., Roque, A., Vaswani, A., Traum, D., 2006. Evaluation of a spoken dialogue system for virtual reality call for fire training. In: Proc. of the 25th Army Science Conference. Orlando, USA.

Roque, A., Ai, H., Traum, D., 2006a. Evaluation of an information state-based dialogue manager. In: Proc. of the 10th Workshop on the Semantics and Pragmatics of Dialogue (Brandial'06). Potsdam, Germany, pp. 181–182.

Roque, A., Leuski, A., Rangarajan, V., Robinson, S., Vaswani, A., Narayanan, S., Traum, D., 2006b. Radiobot-CFF: A Spoken Dialogue System for Military Training. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pp. 477–480.

Rotaru, M., Litman, D. J., 2006. Discourse structure and speech recognition problems. In: Proc. of the 9th Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pp. 53–56.

Rudnicky, A., Thayer, E., Constantinides, P., Tchou, C., Shern, R., Lenzo, K., Xu, W., Oh, A., 1999. Creating natural dialogs in the Carnegie Mellon Communicator system. In: Proc. of European Conference on Speech Communications and Technology (Eurospeech'99). Vol. 1(4). pp. 1531–1534.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., 1986. PDP: Computational models of cognition and perception, I. MIT Press.

Russell, J. A., 1980. A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178.

Sanderman, A., Sturm, J., den Os, E., Boves, L., Cremers, A., 1998. Evaluation of the Dutch train timetable information system developed in the ARISE project. In: Proc. of IEEE 4th Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA'98). Torino, Italy, pp. 91–96.

Sarukkai, R., Hunter, C., 1997. Integration of eye fixation information with speech recognition systems. In: Proc. of the 5th European Conference on Speech Communication and Technology (Eurospeech'97). Rhodes, Greece, pp. 1639–1643.

Schalkwyck, J., Story, L. H. E., 2003. Speech recognition with Dynamic Grammars Using Finite-State Transducer. In: Proc. of Eurospeech'03. Geneva, Switzerland, pp. 1969–1972.

Scherer, K. R., 2005. What are emotions? and how can they be measured? Social Science Information 44 (4), 694–729.

Schiel, F., 2006. Evaluation of multimodal dialogue systems. In: Wahlster (2006), pp. 617–643.

Schneider, M., 2004. Towards a Transparent Proactive User Interface for a Shopping Assistant. In: Proc. of Workshop on Multi-User and Ubiquitous User Interfaces (MU3I). Vol. 3. Funchal, Madeira, Portugal, pp. 10–15.

Schultz, T., Kirchhoff, K., 2006. Multilingual Speech Processing. Elsevier.

Schultz, T., Waibel, A., 1998. Language independent and language adaptive large vocabulary speech recognition. In: Proc. of the 5th International Conference of Spoken Language Processing (ICSLP'98). Vol. 5. Sidney, Australia, pp. 1819–1822.

Scott, W., 1955. Reliability of content analysis: the case of nominal scale coding. Public opinion quarterly 19 (3), 321–325.

Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., Zue, V., 1998. Galaxy-ll: A Reference Architecture for Conversational System Development. In: Proc. of the 5th International Conference on Spoken Language Processsing (ICSLP'98). Vol. 3. Sydney, Australia, pp. 931–934.

Seneff, S., Polifroni, J., 2000. Dialogue management in the Mercury flight reservation system. In: Proc. of ANLP-NAACL Workshop on Conversational systems. Vol. 3. Seattle, Washington, USA, pp. 11–16.

Shafran, I., Mohri, M., 2005. A comparison of classifiers for detecting emotion from speech. In: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05). Philadelphia, PA, USA, pp. 341–344.

Shafran, I., Riley, M., Mohri, M., 2003. Voice signatures. In: Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03). St. Thomas, U.S. Virgin Islands, USA, pp. 31–36.

Sim, J., Wright, C. C., 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy 85 (3), 257–268.

Skantze, G., Edlund, J., Carlson, R., 2006. Talking with Higgins: Research challenges in a spoken dialogue system. In: Proc. of Perception and Interactive Technologies (PIT'06). Kloster Irsee, Germany, pp. 193–196.

Stern, R., Liu, F., Ohshima, Y., Sullivan, T., Acero, A., 1992. Multiple approaches to robust speech recognition.

Stewart, J., 1922. An elecrical analog of the vocal tract. Nature 110, 311–312.

Stibbard, R., 2000. Automated extraction of tobi annotation data from the reading/leeds emotional speech corpus. In: Proc. of the International Speech Communication Association (ISCA) Workshop on Speech and Emotion. Newcastle, Northern Ireland, UK, pp. 60–65.

Sturm, J., Cranen, B., Terken, J., Bakx, I., 2005. Effects of prolonged use on the usability of a multimodal form-filling interface. In: Minker, W., Bühler, D., Dybkjaer, L. (Eds.), Spoken multimodal human-computer dialogue in mobile environments. Springer, pp. 329–348.

Sun, D., 1997. Statistical modeling of co-articulation in continuous speech based on data driven interpolation. In: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97). Munich, Germany, pp. 1751–1754.

The Safira Project - DFKI Page, 2002. `http://www2.dfki.de/imedia/safira/`.

Traum, D., Larsson, S., 2003. The Information State Approach to Dialogue Management. Current and New Directions in Discourse and Dialogue. Kluwer Academic Publishers.

TRINDIConsortium, 2001. TRINDI (Task Oriented Instructional Dialogue) Book Draft. http://www.ling.gu.se/projekt/trindi/book.ps .

Truillet, P., Grisvard, O., Goujon, B., 2004. SCOPE - CARE II Innovative WP3 -R3 - Model of English. Tech. rep., European Organisation for the Safety of Air Navigation.

Turing, A., 1950. Computing machinery and intelligence. Mind 236, 433–460.

Turunen, M., Hakulinen, J., Kainulainen, A., 2006. Evaluation of a spoken dialogue system with usability tests and long-term pilot studies: Similarities and differences. In: Proc. of the 9th Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pp. 1057–1060.

Turunen, M., Salonen, E., Hartikainen, M., Hakulinen, J., Black, W., Ramsay, A., Funk, A., Conroy, A., Thompson, P., Stairmand, M., Jokinen, K., Rissanen, J., Kanto, K., Kerminen, A., Gamback, B., Cheadle, M., Olsson, F., Sahlgren, M., 2004. Athosmail: a multilingual adaptive spoken dialogue system for the e-mail domain. In: Proc. of Workshop on Robust and Adaptive Information Processing for Mobile Speech Interfaces. Geneva, Switzerland, pp. 77–86.

Vaquero, C., Saz, O., Lleida, E., Marcos, J., Canalís, C., 2006. VOCALIZA: An application for computer-aided speech therapy in spanish language. In: Proc. of IV Jornadas en Tecnología del Habla. Zaragoza, Spain, pp. 321–326.

Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: resources, features and methods. Speech communication 48, 1162–1181.

VICTEC in Lynne Hall web page, 2005. `http://osiris.sunderland.ac.uk/~cs0lha/Research/victec.html`.

Vidrascu, L., Devillers, L., 2005. Real-life emotion representation and detection in call centers data. Lecture Notes on Computer Science 3784, 739–746.

Villing, J., Larsson, S., 2006. Dico - A Multimodal Menu-based In-vehicle Dialogue System. In: Proc. of the 10th Workshop on the Semantics and Pragmatics of Dialogue (BRANDIAL'06). Potsdam, Germany, pp. 187–188.

Vogt, T., André, E., 2005. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: Proc. of IEEE International Conference on Multimedia and Expo. pp. 474–477.

Wahlster, W. (Ed.), 2006. SmartKom: Foundations of Multimodal Dialogue Systems. Springer.

Waibel, A., Suhm, B., Vo, M., Yang, J., 1997. Multimodal Interfaces for Multimedia Information Agents. In: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97). Munich, Germany, pp. 167–170.

Walker, M., Cahn, J., Whittaker, S., 1997. Improvising linguistic style: Social and affective bases of agent personality. In: Proc. of the 1st International Conference on Autonomous Agents (Agents'97). Marina del Rey, CA, USA, pp. 96–105.

Walker, M., Fromer, J., Fabbrizio, G., Mestel, C., Hindle, D., 1998a. What can I say? Evaluating a spoken language interface to Email. In: Proc. of ACM CHI 98 Conference on Human Factors in Computing Systems. Los Angeles, USA, pp. 582–589.

Walker, M., Kamm, C. A., Litman, D. J., 2000a. Towards developing general models of usability with paradise. Natural Language Engineering, 363–377.

Walker, M., Langkilde, I., wright, J., Gorin, A., Litman, D., 2000b. Learning to predict problematic situations in a spoken dialogue system: Experiments with How May I Help You. In: Proc. of the North American Meeting of the Association for Computational Linguistics. Seattle, USA, pp. 210–217.

Walker, M., Litman, D., Kamm, C., Abella, A., 1998b. Evaluating spoken dialogue agents with PARADISE: Two case studies. Computer Speech and Language 12, 317–347.

Walker, M., Rudnicky, A., Prasad, R., Aberdeen, J., Bratt, E. O., Garofolo, J., Hastie, H., Le, A., Pellom, B., Potamianos, A., Passonneau, R., Roukos, S., Sanders, G., Seneff, S., Stallard, D., 2002a. DARPA Communicator: Cross-System Results for The 2001 Evaluation. In: Proc. of the 7th International Conference on Spoken Language Processing (Interspeech'02-ICSLP). Vol. 1. Denver, USA, pp. 269–272.

Walker, M., Rudnicky, A., Prasad, R., Aberdeen, J., Bratt, E. O., Garofolo, J., Hastie, H., Le, A., Pellom, B., Potamianos, A., Passonneau, R., Roukos, S., Sanders, G., Seneff, S., Stallard, D., 2002b. DARPA Communicator Evaluation: Progress from 2000 to 2001. In: Proc. of the 7th

International Conference on Spoken Language Processing (Interspeech'02-ICSLP). Vol. 1. Denver, USA, pp. 273–276.

Weizenbaum, J., 1966. ELIZA - A computer program for the study of natural language communication between man and machine. Communications of the ACM 9, 36–45.

Weng, F., Varges, S., Raghunathan, B., Ratiu, F., Pon-Barry, H., Lathrop, B., Zhang, Q., Scheideck, T., Bratt, H., Xu, K., Purver, M., Mishra, R., Raya, M., Peters, S., Meng, Y., Cavedon, L., Shriberg, L., 2006. CHAT: A Conversational Helper for Automotive Tasks. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pp. 1061–1064.

Wilks, Y., 2006. Artificial companions as a new kind of interface to the future internet. Tech. Rep. 13, Oxford Internet Institute.

Williams, J., Young, S., 2007. Partially Observable Markov Decision Processes for Spoken Dialog Systems. Computer Speech and Language 21(2), 393–422.

Wilting, J., Krahmer, E., Swerts, M., 2006. Real vs. acted emotional speech. In: Proc. of the 9th Conference on Spoken Language Processing (Interspeech-ICSLP). Pittsburgh, USA, pp. 805–808.

Witten, I. H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Zeng, Z., Hu, Y., Fu, Y., Huang, T. S., Roisman, G. I., Wen, Z., 2006. Audio-visual emotion recognition in adult attachment interview. In: Proc. of the 8th International Conference on Multimodal interfaces. Banff, Alberta, Canada, pp. 828–831.

Zgank, A., Kaèiè, Z., Diehl, F., Vicsi, K., Szaszak, G., Juhar, J., Lihan, S., 2004. The COST278 MASPER initiative - Crosslingual speech recognition with large telephone databases. In: Proc. of the 4th International Conference on Language Resources and Evaluation (LREC 2004). Lisbon, Portugal, pp. 2107–2110.

Zong, Y., Dohi, H., Ishizuka, M., 2000. Multimodal presentation markup language mpml with emotion expression functions attached. In: Proc. of the 2nd International Symposium on Multimedia Software Engineering. pp. 359–365.

Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., Hetherington, L., 2000. JUPITER: A telephone-based conversational interface for weather information. In: IEEE Transactions on Speech and Audio Processing. Vol. 8. pp. 85–96.

Zue, V. W., Glass, J. R., 2000. Conversational interfaces: Advances and challenges. Proc. of the IEEE 88, 1166–1180.

The Thesis presents the work done on three of the most challenging topics in the area of spoken dialogue systems: recognition of non-acted emotions, cross-lingual adaptation of speech recognizers and field evaluation. The research described constitutes a novel contribution to what the experts have established as the major research trends in the area of SDS: adaptiveness and portability of the systems.

Firstly, regarding emotion recognition, a detailed study is supplied on how to calculate and interpret reliability coefficients for the annotation of corpora of real emotions. A new efficient approach is proposed that considerably enhances inter-annotator agreement and machine emotion recognition by the use of several context information sources.

Secondly, the research on cross-lingual adaptation of speech recognizers was carried out during a three-month stay at the Technical University of Liberec (Czech Republic). An approach is presented to cost-efficiently (in terms of time and effort) adapt a speech recognizer to work in another language. The proposal has been used to adapt a Czech speech recognizer to a language which is acoustically very similar (Slovak) and another with a completely different origin (Spanish).

Thirdly, several statistical studies were carried out on a field evaluation of a spoken dialogue system. New empirical evidence is provided on the relationships between evaluation criteria. The study includes both interaction parameters and quality judgments, paying special attention to user satisfaction and task success.

All the methods proposed in the Thesis have been tested with real dialogue systems, for which the UAH spoken dialogue system was developed.