LAS TECNOLOGÍAS DEL HABLA: ENTRE LA INGENIERÍA Y LA LINGÜÍSTICA

JOAQUIM LLISTERRI
Departament de Filologia Espanyola
Universitat Aut noma de Barcelona
Joaquim.Llisterri@uab.es
http://liceu.uab.es/~joaquim/home.html

1.- Introducción

La visión de las tecnologías del habla que tiene el gran público suele girar en torno a las máquinas que hablan y entienden, cuyo referente desde 1968 ha sido el Hal que protagonizara *2001: una odisea del espacio*. Más recientemente, las voces electrónicas se han incorporado a diversos servicios telefónicos de uso común, de modo que cada vez resulta menos extraño encontrar un ordenador que responda a una llamada substituyendo a los habituales operadores (Llisterri, 2001a).

Si preguntáramos a los usuarios de estos servicios qué tipo de profesionales imagina que los desarrollan, dirían casi con toda seguridad que ingenieros, informáticos, tecnólogos o científicos; en muy pocos casos una persona no especializada en el campo pensaría en la lingüística como una de las disciplinas que hacen posible que un ordenador lea en voz alta o conteste a nuestras preguntas. Incluso entre los propios profesionales de la lingüística y de la filología, no es infrecuente creer que los recientes portales de voz, por ejemplo, son ajenos a su labor habitual.

Desde una perspectiva histórica, no deja de ser cierto que las tecnologías del habla nacieron de la mano de la ingeniería de las telecomunicaciones, motivadas por la necesidad de reducir, mediante la codificación de la voz, la cantidad de información que se transmitía a través del hilo telefónico. Sin embargo, la propia denominación de la disciplina algo indica sobre la dualidad del campo de trabajo, puesto que si, por una parte, existe un indudable componente tecnológico, por otra, el objeto que se manipula es el habla, medio de comunicación humana cuyo estudio es una de las tareas de la lingüística.

Hoy en día, las tecnologías del habla siguen siendo un terreno estrechamente vinculado a la ingeniería ya que, en última instancia, se persigue el desarrollo de aplicaciones reales. Sin embargo, muchos centros de investigación y desarrollo han comprendido que las mejoras en sus sistemas pueden depender no sólo de técnicas más avanzadas de programación o de algoritmos más potentes de tratamiento de señales, sino también de un buen conocimiento de los mecanismos que subyacen a la producción y la percepción del habla.

La necesidad de incorporar conocimientos propios de la lingüística a las tecnologías del habla ha hecho que cada vez sea más frecuente la participación de lingüistas – especialmente de los que centran su actividad en la fonética, rama dedicada al estudio del habla – en proyectos

conjuntos con grupos de ingenieros e informáticos. El presente trabajo pretende, precisamente, ilustrar el papel de un equipo de especialistas en fonética en el desarrollo de diversos sistemas en el ámbito de las tecnologías del habla, presentando algunos de los trabajos llevados a cabo por el Grupo de Fonética del Departamento de Filología Española de la Universidad Autónoma de Barcelona (UAB)[1].

2.- Conversión de texto en habla

La conversión de texto en habla es una tecnología que permite llevar a cabo la lectura en voz alta por parte de un sistema informático de cualquier texto almacenado en formato electrónico (Llisterri, 2001b). Se utiliza, por ejemplo, en los portales de voz que proporcionan información que, por su naturaleza cambiante, no es práctico que la grabe previamente un locutor humano, como puede ser la del tiempo o del tráfico, o en la lectura del correo electrónico a través del teléfono, puesto que el texto que recibirá un usuario no se puede predecir. La conversión de texto en habla es también la tecnología que emplean los invidentes para tener acceso a la información contenida en una página web, a un documento escrito con un procesador de textos, o a un mensaje de correo.

En un sistema de conversión de texto en habla existen módulos dedicados al procesamiento de los datos lingüísticos que requieren para su desarrollo la intervención de un equipo especializado en este campo. Precisamente por ello, el Grupo de Fonética de la UAB ha intervenido en las versiones en español de los conversores desarrollados por el CNET *Centre National d'Etudes des Télécommunications* (Lannion, Francia)[2] entre 1992 y 1996, y por el CSELT *Centro Studi e Laboratori Telecomunicazioni* (Turín, Italia)[3] entre 1998 y 2000, así como en la creación de la versión en catalán de los conversores de Telefónica I+D (Madrid) entre 1995 y 1999, de Loquendo (Turín, Italia) desde principios de 2002 y del CREL *Centre de Referència en Enginyeria Lingüística de la Generalitat de Catalunya* entre 1996 y 2000[4].

2.1.- TRATAMIENTO PREVIO DEL TEXTO

La primera tarea a la que se enfrenta el lingüista que participa en el desarrollo de un conversor de texto en habla es el tratamiento previo del texto que, posteriormente, se transformará en su correspondiente realización sonora. Debe tenerse en cuenta que un conversor exige, para su buen funcionamiento, que todos los elementos se presenten como una cadena de caracteres que pueda ser leída. Por tal motivo, las abreviaturas – tratamientos y monedas, por ejemplo -, las

---

[1] La información sobre el grupo se encuentra en `http://liceu.uab.es`. Algunos de los trabajos y de las líneas de investigación se resumen en Aguilar *et al*. (1997), Llisterri (1997) y Llisterri *et al.* (1999).

[2] Un ejemplo de este trabajo es la voz española "Rafael" del sistema de conversión de texto en habla comercializado por la empresa francesa Elan (`http://www.elantts.com/`).

[3] Las voces españolas " Juan" y "Carmen" y la catalana "Montserrat" del sistema Actor comercializado por la empresa italiana Loquendo (`http://www.loquendo.com/`) son el resultado de esta colaboración.

[4] Puede accederse a demostraciones de este sistema, desarrollado conjuntamente con el *Grup de Tractament de la Parla* de la Universidad Politécnica de Cataluña, en `http://gps-tsc.upc.es/veu/`. Parte del conversor en catalán comercializado por la empresa Atlas (`http://www.atlas-cti.com/`) se basa también en módulos de procesamiento lingüístico creados en el marco del CREL.

siglas y los acrónimos, los números arábigos y romanos, fechas, horas, los símbolos especiales como el del euro o el dólar, etc. deben transformarse en sus equivalentes deletreados. Eso supone la confección de listas en las que cada uno de estos elementos está asociado a su correspondiente expresión completa y, en algunos casos, a una transcripción fonética.

Aunque esta pueda parecer una labor exenta de dificultades, es preciso considerar, por ejemplo, alternancias en siglas como PSOE, pronunciada a veces como [pe'soe], como ['soe] o como ['psoe], y prever aquellos casos en que una sigla o un acrónimo pueden transcribirse automáticamente (RENFE, ONU) como si se trataran de cualquier otra palabra o en que deben deletrearse (ADSL, SMS). También en los números hay que prever un conjunto de reglas de concordancia para que, pongamos por caso, 300 se transforme en "trescientos" cuando va acompañado de un nombre masculino y en "trescientas" cuando le sigue uno femenino. Lo mismo sucede en los símbolos de las monedas, que deben concordar en singular o en plural según si delante aparece 1 u otra cantidad.

2.2.- ANÁLISIS LINGÜÍSTICO

Otro de los módulos necesarios para la conversión de texto en habla es que lleva a cabo el análisis lingüístico del texto; en el caso de algunos de los conversores antes mencionados, el Grupo de Fonética de la UAB ha desarrollado las reglas de los programas conocidos como "categorizadores", que asignan automáticamente a cada palabra del texto la información sobre la parte de la oración o categoría gramatical a la que pertenecen y, en algunos casos, otra información morfológica. El desarrollo de un categorizador implica, por tanto, un conocimiento de las propiedades formales de las palabras adaptado a los requisitos de un sistema que debe operar de un modo automático. Esta información es crucial, por ejemplo, para evitar una pausa entre un nombre y el artículo que le precede o para determinar la entonación de una frase a partir de los elementos interrogativos que en ella aparecen.

2.3.- TRANSCRIPCIÓN FONÉTICA AUTOMÁTICA

Transformar un texto en su equivalente oral requiere un paso intermedio: la generación de una representación o transcripción fonética, con el fin de acercar la forma ortográfica de las palabras a su manifestación sonora. Esta operación se realiza mediante un conjunto de reglas que asignan automáticamente a cada carácter ortográfico su correspondiente realización fonética. En español, por ejemplo, <h> debe eliminarse excepto cuando forma parte del dígrafo <ch> y <j> y <g> deben asociarse al mismo sonido [x] cuando <j> va seguida de <a, o, u> y cuando <g> precede a <e, i>. Estas mismas reglas tienen también que dar cuenta de la sílaba acentuada de la palabra cuando esta se escribe sin tilde, evitando, sin embargo, la colocación de acento en artículos, ciertas formas auxiliares de los verbos – "ha", pero no "habrá" – o determinadas preposiciones – "a", "de", etc. -.

Puede deducirse fácilmente que para el diseño de este tipo de reglas se necesita una estrecha familiaridad con la descripción fonética de la lengua y con la norma que rige la pronunciación estándar en los casos conflictivos. Una situación especialmente compleja es la que se produce en la transcripción de topónimos y antropónimos de otras lenguas, para los que muchas veces no existe una forma normativamente establecida. Una transcripción que responda a una pronunciación completamente nativa supondría introducir en el conversor unidades de síntesis propias de otras lenguas, como se hace, por ejemplo, con [T], que se incorpora a los sistemas en catalán aunque no sea un sonido de esta lengua para la pronunciación de apellidos castellanos

acabados en <z>. Sin embargo, una pronunciación totalmente nativa de un nombre japonés, ruso, alemán o incluso inglés podría presentar problemas a los usuarios no familiarizados con estas lenguas. Por ello suele llegarse habitualmente a una solución de compromiso, al igual que ocurre en los medios de comunicación orales, "castellanizando" o "catalanizando", según la lengua del conversor, los nombres propios.

Otras decisiones que se toman en el momento de diseñar un sistema de transcripción fonética automática responden, en el fondo, a la elección de un modelo de pronunciación. Por ejemplo, en español es preciso optar por una de las múltiples realizaciones de los participios acabados en <ado> o de las palabras acabadas en <d>, así como establecer la realización fonética de <ll> como [j] o como [¥].

La transcripción fonética automática es, pues, una operación que exige conocimiento lingüístico y, por esta razón, es uno de los campos en los que más ha trabajado el Grupo de Fonética de la UAB, tanto para el castellano (Ríos, 1993, 1999) como para el catalán en los proyectos anteriormente mencionados. Cabe destacar el sistema Segre desarrollado para el catalán (Pachès *et al*., 2000; de la Mota y Riera, 2000) en colaboración con el *Grup de Tractament de Parla* de la Universidad Politécnica de Cataluña. Segre es una herramienta que transcribe automáticamente textos en catalán, en cuyo diseño se tuvo en cuenta que las reglas fueran fácilmente modificables por expertos con formación lingüística, ya que utilizan un formalismo similar al usado en fonología y, además, fueran relativamente independientes del programa, de modo que pudiera comprobarse el efecto de una regla o del orden en el que se aplican sin necesidad de alterar el algoritmo.

2.4.- DICCIONARIOS DE UNIDADES DE SÍNTESIS

Los conversores de texto en habla se basan, por lo general, en un diccionario de unidades de síntesis a partir de las que se construyen los enunciados. Estas unidades – denominadas en algunos sistemas "difonemas"[5] –, tras su grabación por un locutor y su codificación en una serie de parámetros acústicos, se almacenan en un diccionario del que se seleccionan para concatenarlas y formar mensajes en función de la transcripción fonética del texto de entrada. Muchas de las etapas del proceso de creación de un diccionario de unidades de síntesis requieren decisiones de tipo lingüístico, tal como se expone a continuación, reflejando las experiencias en el desarrollo de diccionarios en castellano y en catalán del Grupo de Fonética de la UAB.

El primer paso para construir un diccionario de unidades de síntesis es definir el inventario de alófonos que utilizará el conversor. Esto requiere un estudio de la totalidad de los alófonos que configuran el sistema fonético de la lengua y la toma de decisiones sobre los que, debido a su similitud acústica o perceptiva, pueden excluirse sin que resulte afectada la calidad de la síntesis. A continuación, se establecen las posibles combinaciones entre alófonos, recurriendo para ello al conocimiento sobre la estructura silábica de la lengua y a las restricciones fonotácticas, que establecen qué elementos pueden aparecer o no en una determinada posición.

[5] La propia denominación de las unidades es una muestra más del origen histórico de las tecnologías del habla al que aludíamos al principio. Se trata de unidades compuestas por dos segmentos sonoros que corresponden a lo que en fonética y en fonología se definen como "alófonos" o "fonos", y no como "fonemas", por lo que el término adecuado sería realmente "dialófono".

Una vez determinadas las unidades, se selecciona el locutor que llevará a cabo su grabación – en general incluidas en frases o en textos, pues la lectura de difonemas aislados sería muy poco natural – y que constituirá la voz del sistema de conversión. Intervienen aquí de nuevo criterios lingüísticos, pues es importante verificar que la persona elegida tenga un acento que corresponda al estándar deseado para el conversor, además de una voz que ofrezca buenos resultados cuando sea manipulada para la síntesis. Durante el proceso de grabación es necesaria la presencia de un experto en fonética para asegurar la adecuada realización de cada una de las unidades según el inventario establecido así como una cierta homogeneidad en el ritmo y la entonación, ya que posteriormente la síntesis se realizará con unidades recogidas en distintas sesiones de grabación y procedentes de enunciados diferentes.

En algunas ocasiones, antes de proceder a la grabación del corpus de unidades completo – tarea que puede llevar varios días en un estudio - se realizan pruebas con un corpus reducido y con varios locutores. El objetivo es conseguir que un número lo más representativo posible de futuros usuarios del sistema de conversión escuchen un conjunto de enunciados con diferentes voces sintetizadas, de modo que los resultados puedan tenerse en cuenta en la selección definitiva del locutor. El diseño de estas pruebas subjetivas de evaluación sigue, por lo general, el de los experimentos en fonética perceptiva, por lo que la experiencia en este campo representa una contribución relevante al proyecto.

Existen hoy en día técnicas de conversión de texto en habla que se basan en grandes corpus – es decir, conjuntos estructurados de textos – de los que se extraen las unidades necesarias para la síntesis. La tarea del lingüista en estos casos es seleccionar del corpus aquellos enunciados que se incorporarán, una vez grabados, a la base de datos del sistema, intentando combinar criterios estadísticos con la necesaria adecuación gramatical de los fragmentos elegidos.

2.5.- MODELOS PROSÓDICOS PARA LA CONVERSIÓN DE TEXTO EN HABLA

Uno de los aspectos que contribuye a dotar de una mayor naturalidad a un sistema de conversión de texto en habla y, por tanto, a favorecer su aceptación por parte de sus potenciales usuarios, es sin duda la prosodia. En la lectura de un texto se varía la duración de los sonidos, la intensidad con la que se pronuncia cada uno de ellos, la entonación de las frases, la velocidad de elocución, y el ritmo; también es un ingrediente esencial para una buena lectura la acertada colocación de las pausas, que contribuye a estructurar el texto y a aumentar la facilidad para comprenderlo. Todos estos aspectos, buena parte de ellos de tipo fonético, deben reflejarse en las reglas de un conversor que se encargan de la asignación de prosodia al texto de entrada.

El problema principal que se aborda en esta etapa es la falta de información en un texto escrito sobre sus características prosódicas en el momento de leerlo. Las pausas, por ejemplo, vienen marcadas por los signos de puntuación, pero todo buen lector realiza pausas que en ocasiones no tienen una representación ortográfica y que se establecen en función del sentido; otras veces, como en el caso de los correos electrónicos escritos de manera rápida e informal, los signos de puntuación pueden ser escasos o estar situados en puntos que no faciliten la comprensión del mensaje. La entonación tiene también su representación ortográfica en los signos de interrogación y de exclamación, aunque la situación es algo más compleja, puesto que una coma puede requerir una inflexión tonal, que será distinta en una enumeración o en un inciso. Incluso la entonación de una pregunta será diferente en función de la presencia o ausencia de un pronombre interrogativo, o de otros factores como la intencionalidad de quien escribió el texto.

Un texto escrito tampoco contiene información sobre la duración o la intensidad de cada sonido, mientras que los estudios en fonética experimental han demostrado que éstas pueden verse modificadas por el acento, por la posición del segmento en el enunciado o por la presencia de pausas, entre otros factores. Por ello es preciso definir una serie de reglas que determinen en cada caso, partiendo de unos valores intrínsecos de base, las modificaciones necesarias si se desea alcanzar una buena conversión en habla. Para esta tarea es imprescindible disponer de un corpus de textos leídos, cuidadosamente diseñado de modo que se contemplen todos los posibles factores de variación (Riera y Jiménez, 2000).

El Grupo de Fonética de la UAB ha centrado buena parte de su dedicación en la obtención de la información lingüística necesaria para el módulo prosódico de los conversores de texto en habla mencionados en el apartado 1, tanto para el castellano como para el catalán (Llisterri *et al*., 2000). Además, en el marco de la colaboración con la empresa Loquendo (Turín), ha desarrollado modelos de entonación específicamente para las oraciones interrogativas que se han aplicado al español de México, el portugués de Brasil, el inglés y el alemán, siguiendo la metodología expuesta en Garrido *et al*. (2000).

Los trabajos llevados a cabo se han basado, en algunos casos, en investigaciones previas sobre la entonación en español orientadas a su aplicación en la síntesis (Garrido, 1991, 1996, 2001) y en estudios encaminados a determinar los factores que inciden en la duración (Marín, 1994) y la intensidad (Blecua y Acín, 1995) de las vocales o la adecuada asignación de las pausas en un texto (Puigví *et al*., 1994). También para el catalán se han analizado las propiedades acústicas de los patrones melódicos (Estruch, 2000) y del grupo acentual (Riera, 1999, 2001), con objeto de incorporar los resultados a las reglas prosódicas de un conversor de texto en habla.

2.6.- EVALUACIÓN DE SISTEMAS DE CONVERSIÓN DE TEXTO EN HABLA

Una de las necesidades que surgen en cuanto se desarrolla un conversor de texto en habla es disponer de un procedimiento de evaluación que permita comprobar, desde el punto de vista del usuario, las mejoras realizadas en las sucesivas versiones del sistema. Por este motivo, el Grupo de Fonética de la UAB colaboró con Telefónica I+D (Madrid) entre 1993 y 1999 en la creación de un conjunto estandarizado de pruebas de evaluación y diagnóstico para conversores de texto en habla en castellano y en catalán (Aguilar *et al*., 1994 a, b).

En las pruebas se evalúan básicamente tres aspectos: la inteligibilidad, la comprensión y la calidad global. En lo que se refiere a la inteligibilidad, se utilizan pruebas de base fonética en las que se determina la identificación de consonantes, de grupos consonánticos y de vocales en contacto, además del reconocimiento de palabras en frases con sentido y sin sentido. La comprensión se valora mediante la respuesta a una serie de preguntas tras escuchar un texto leído por el conversor, mientras que la calidad global se cuantifica a partir de la respuesta en una escala del 1 al 5 a una serie de preguntas orientadas a descubrir la aceptación de posibles servicios por parte de un grupo representativo de futuros usuarios.

Es también interesante señalar que pueden igualmente diseñarse procedimientos de evaluación para usuarios específicos; en este sentido, el Grupo de Fonética de la UAB llevó a cabo, en colaboración con la Universidad Ramon Llull, un estudio sobre el uso del conversor de texto en habla Ciber232 por parte de invidentes (Llisterri *et al*., 1993).

3.- Reconocimiento del habla

El reconocimiento puede entenderse como una tecnología que realiza la misma operación que la conversión de texto en habla pero en sentido inverso; se trata, en efecto, de transformar una señal sonora – es decir, el habla – en su equivalente escrito. La aplicación más conocida es, seguramente, el dictado automático, que cuenta con productos comerciales orientados a profesionales de la medicina, la jurisprudencia, la traducción, el periodismo, y a todas aquellas personas que prefieren utilizar la voz en lugar del teclado para redactar sus documentos.

El Grupo de Fonética de la UAB participó, conjuntamente con el *Grup de Tractament de la Parla* de la Universidad Politécnica de Cataluña, en el desarrollo de la versión catalana del sistema de dictado automático *FreeSpeech*, comercializado por la empresa Philips en 1999[6], así como en la elaboración del corpus en catalán *SpeechDat*, utilizado para entrenar y evaluar programas de reconocimiento de habla en esta lengua[7]. En lo que se refiere al español, el Grupo formó parte del consorcio *Albayzín*, que dio lugar al corpus del mismo nombre (Casacuberta *et al*., 1992; Moreno *et al*., 1993), especialmente diseñado para la creación de aplicaciones en reconocimiento del habla y actualmente a disposición de investigadores y empresas a través de la Agencia Europea de Distribución de Recursos Lingüísticos (ELDA).

Una de las principales tareas en las que interviene un equipo de lingüistas implicado en el desarrollo de un sistema de reconocimiento es el asesoramiento sobre la muestra de locutores seleccionada para entrenar el sistema. Los reconocedores de habla se basan, en términos generales, en "plantillas" o modelos de cada uno de los sonidos y combinaciones de sonidos de la lengua; estos modelos se determinan mediante técnicas estadísticas de aprendizaje automático aplicadas al análisis de extensos corpus que datos que deben incorporar la mayor variedad posible de hablantes. Por este motivo es preciso considerar las variantes geográficas propias de cada lengua y establecer un procedimiento de recogida de datos que las refleje adecuadamente.

El proceso de aprendizaje requiere que el corpus se encuentre adecuadamente segmentado y etiquetado, es decir, que a la señal sonora grabada se incorporen marcas que indiquen el principio y el final de cada sonido así como la correspondiente transcripción fonética, sincronizada con la representación ortográfica. El conocimiento de la fonética acústica es crucial en esta fase para establecer criterios homogéneos de segmentación, de etiquetado y de transcripción y para, en su caso, revisar manualmente los errores que inevitablemente se producen cuando se aplican programas informáticos que realizan automáticamente estas operaciones.

Un reconocedor suele contener un diccionario en el que se encuentran fonéticamente transcritas las palabras que aceptará el sistema. Es una labor propia del lingüista definir no únicamente la pronunciación "canónica" de cada palabra, sino también aquellas variantes que se han

---

[6] El proyecto contó con financiación de la *Generalitat de Catalunya*. Recientemente Philips dejó de comercializar este producto en todas las lenguas.

[7] Parte del corpus *SpeechDat* en catalán se desarrolló en el marco del CREL *Centre de Referència en Enginyeria Lingüística de la Generalitat de Catalunya* siguiendo los estándares del proyecto europeo *SpeechDat* (`http://www.speechdat.org/`). Puede encontrase más información sobre el corpus catalán en `http://gps-tsc.upc.es/`.

encontrado en el corpus de entrenamiento al que antes aludíamos y, si es preciso, establecer reglas que definan la correspondencia entre ambas.

Todo ello muestra que, al igual que la conversión, el reconocimiento del habla es una tecnología a la que, en la práctica, se incorpora conocimiento lingüístico especializado.

4.- Sistemas de diálogo

Gran parte de los servicios basados en las tecnologías del habla que están alcanzando actualmente mayor popularidad consisten en los denominados "sistemas de diálogo", a través de los cuales un usuario establece una interacción oral con un ordenador, sea para obtener una información o para realizar una determinada transacción. Los portales de voz mencionados al principio constituyen tal vez el ejemplo más claro de las tecnologías a las que nos estamos refiriendo, aunque existen otras posibilidades como la banca electrónica, los sistemas de cita previa o los servicios automáticos de atención al cliente.

Un sistema de diálogo incorpora un reconocedor para procesar las preguntas del usuario y un conversor con el fin de proporcionar oralmente las respuestas; el módulo central, en cambio, lo constituye el llamado "gestor del diálogo" que establece los turnos de palabra, se encarga de mantener la coherencia entre la pregunta y la respuesta, interpreta las intervenciones del usuario que hacen referencia a información previa o a elementos deícticos – por ejemplo "mañana", "aquí", etc. – y, en conjunto, pone en práctica la estrategia diseñada por los investigadores para que la interacción se lleve a cabo del mejor modo posible.

La realización de un sistema de diálogo suele iniciarse con el estudio de intercambios comunicativos reales entre personas o con el análisis de interacciones simuladas entre personas y máquinas; esto lleva a establecer un conjunto de estrategias y a la definición de las intervenciones más adecuadas por parte del ordenador. Nada de ello es ajeno al conocimiento lingüístico, y por tal motivo el Grupo de Fonética de la UAB ha intervenido en el diseño del prototipo de sistema de diálogo desarrollado por el CREL *Centre de Referència en Enginyeria Lingüística de la Generalitat de Catalunya* entre 1996 y 2000, y en la puesta en marcha a mediados de 2002 del sistema de información meteorológica por teléfono en catalán *aTTemps*[8], promovido conjuntamente por el Departamento de Medio Ambiente y el Departamento de Universidades, Investigación y Sociedad de la Información de la *Generalitat de Catalunya*.

4.1.- Corpus para el desarrollo de sistemas de diálogo

Mencionábamos anteriormente que el diseño de un sistema de diálogo parte, en general, del estudio de la interacción entre personas que realizan la tarea que se pretende automatizar. En el caso del sistema del CREL, se recogió un corpus de llamadas a los servicios de información de los *Ferrocarrils de la Generalitat*, ya que el objetivo era elaborar un prototipo de servicio de información en el caso particular de este medio de transporte público que opera en el territorio catalán. El análisis de este corpus permitió determinar el tipo de consultas realizadas y, a la vez, estudiar la variedad de formas empleadas por los usuarios para solicitar información.

[8] Puede encontrarse información más detallada sobre a*TTemps* en `http://gpstsc.upc.es/veu/attemps/`. El proyecto se desarrolló en colaboración con el *Centre de Tecnologies i Aplicacions del Llenguatge i la Parla* de la Universidad Politécnica de Cataluña.

Sin embargo, es sabido que las personas no actuamos del mismo modo cuando nos dirigimos a un interlocutor humano que en el momento en que nos damos cuenta de que nos atiende un sistema automático. Para obtener un corpus más realista, se recurrió al protocolo conocido como "el Mago de Oz": el usuario que realiza una llamada escucha al otro lado del hilo telefónico una voz sintetizada, pero las respuestas no las genera el sistema de diálogo – puesto que en esta fase del trabajo aún no está operativo -, sino que las teclea un investigador en un conversor de texto en habla, basándose en unos modelos previamente establecidos. Así, se consigue un corpus de interacciones entre persona y máquina que responde con fidelidad al funcionamiento real del sistema que se está diseñando (Machuca *et al*., 2000a).

La necesidad de contar con estos corpus, no únicamente procedentes de llamadas telefónicas, sino también recogidos en entornos caracterizados por la multimodalidad – videoconferencia a través de la web, por ejemplo - ha propiciado el surgimiento de una serie de iniciativas, financiadas en el marco del programa europeo para la Sociedad de la Información, encaminadas a definir estándares para la anotación y codificación de recursos lingüísticos multimodales. El Grupo de Fonética de la UAB ha participado en MATE, *Multilevel Annotation*, *Tools Engineering* entre 1997 y 2000, ISLE-HLT *International Standards for Language Engineerin*g entre 2000 y 2001 y en NITE *Natural Interactivity Tools Engineering*, que se desarrolla entre 2001 y 2003[9]. La contribución del equipo, atendiendo a su especialización, se ha centrado en la anotación fonética – especialmente en lo que se refiere a la entonación (Mengel *et al*., 2000) - y en el establecimiento de relaciones entre elementos gestuales y la prosodia (Dybkjaer *et al*., 2001; Wegener *et al*., 2002).

4.2.- DISEÑO DE ESTRATEGIAS DE DIÁLOGO

El análisis de los corpus que recogen interacciones entre personas y entre potenciales usuarios y sistemas de diálogo simulados permite establecer un conjunto de "escenarios" posibles - los diversos tipos de consultas, por ejemplo - y, para cada uno de ellos, las intervenciones que debe realizar el sistema de diálogo para que quien lo emplea llegue a obtener la información deseada (Machuca *et al*., 2000b). Es evidente para un lingüista que uno de los problemas principales que aquí se plantean es el reconocimiento de los actos de habla: una petición de información, pongamos por caso, no siempre se expresa de forma directa mediante un "Quiero saber..." o con una pregunta del tipo "¿A qué hora hay trenes a...?" o "¿Está lloviendo en...?". Existen ambigüedades como, por ejemplo, "cuándo" referido al día o a la hora, y deben descartarse también aquellas partes de la pregunta que no son relevantes.

Algo esencial de cara al usuario, y también para la imagen pública de la institución o empresa que ofrece el servicio, es la corrección lingüística en las intervenciones del sistema automático. No basta que las preguntas y las respuestas sean precisas, sino que también deben ser pragmáticamente adecuadas – con el suficiente grado de cortesía, por ejemplo – y normativamente aceptables. La revisión de los diálogos por parte de un equipo de expertos familiarizados con estos aspectos del uso lingüístico – tal como se hizo en el proyecto *aTTemps* - parece pues, como mínimo, aconsejable.

---

[9] Más información sobre estos proyectos se recoge en `http://mate.nis.sdu.dk/`, `http://isle.nis.sdu.dk/` y en `http://nite.nis.sdu.dk/`.

Así pues, una vez más encontramos una tecnología que puede mejorar substancialmente con la integración de conocimiento lingüístico, esta vez no tan sólo en el nivel fonético, sino también en otros niveles de descripción de la lengua, como puedan ser el semántico y el pragmático.

5.- Reflexiones finales

Esta presentación ha pretendido poner de manifiesto que el conocimiento lingüístico constituye un elemento indispensable en el desarrollo de las tecnologías del habla y que, en consecuencia, un equipo integrado por expertos formados en el campo tradicionalmente considerado "de Letras" puede jugar un papel relevante en el diseño, el desarrollo y la evaluación de productos y servicios que cada vez estarán más extendidos en la nueva Sociedad de la Información.

Pese a la visión optimista que pueda desprenderse de la lectura de las páginas anteriores, no debe esconderse la existencia de ciertos obstáculos. En primer lugar, la "cultura" dominante en el área de las tecnologías del habla sigue siendo la de los ingenieros, pues al fin y al cabo, éstos son los responsables últimos de transformar una idea en un producto; no todos los equipos universitarios ni todas las empresas son igualmente receptivas a la colaboración con lingüistas, tanto por razones prácticas – aumentan el coste del producto y deben repartirse los beneficios o la financiación de un proyecto – como por motivos derivados de la tradición y de la imagen que las humanidades han proyectado hasta ahora. Por otra parte, el diálogo en los primeros contactos entre ambas comunidades no siempre es fluido, pues existe la idea preconcebida – y en parte cierta – de que los lingüistas están excesivamente interesados en las teorías y los ingenieros demasiado centrados en la aplicación. Incluso la propia terminología (por ejemplo las diferencias comentadas en la nota 3 o el hecho de que unos empleemos "conversión de texto en habla" y "reconocimiento del habla" mientras que otros utilizan "conversión texto-voz" y "reconocimiento de voz"[10]) actúa en ocasiones como un elemento más de desencuentro.

Si bien desde los sectores humanísticos suele reclamarse, muchas veces con razón, un cambio de mentalidad por parte de los tecnólogos, no deja de ser cierto que, en el caso que nos ocupa, es preciso también llegar a comprender el tipo de conocimiento lingüístico que exige el desarrollo de una aplicación. La descripción fonética de la lengua practicada al estilo tradicional no es ciertamente la idónea para incorporarla fácilmente a un conversor de texto en habla, a un reconocedor, o a un sistema de diálogo; el desconocimiento de los principios en los que se basa una determinada tecnología tampoco ayuda a hallar la manera más acertada de obtener e integrar los datos lingüísticos necesarios.

Evidentemente, la posibilidad de que los lingüistas participen en un equipo interdisciplinar conlleva la necesidad de un cambio notable en la formación que éstos reciben. Parece claro que el perfil filológico que domina en la mayoría de los planes de estudio de las facultades de Letras no contribuye, quizás salvo contadísimas excepciones que se traducen en unas pocas asignaturas, a preparar profesionales que puedan intervenir, en pie de igualdad, en el diseño y el desarrollo de los productos que hoy en día ofrecen las tecnologías del habla. Esto tiene como consecuencia que, en ciertos casos, los lingüistas que logran entrar en el mundo tecnológico lo

[10] La voz es, en principio, el resultado de la vibración de las cuerdas vocales; el habla, en cambio, es el medio empleado más comúnmente por los seres humanos para comunicarnos. Por ello, los lingüistas solemos pensar que emitir voz no equivale necesariamente a hablar y que reconocer la voz de una persona no implica interpretar el mensaje que nos quiere transmitir.

hagan reducidos al mero papel de proveedores de datos o de revisores de la información obtenida por procedimientos automáticos, sin que tengan una intervención real en la concepción del proyecto[11].

Cabe pensar, en conclusión, que las tecnologías del habla constituyen un sector idóneo para contribuir a difuminar los fronteras entre la ingeniería y la lingüística, siempre y cuando se olviden antiguos prejuicios y cada especialidad se valore en su justa medida. El mejor modo de acercarse es, seguramente, que cada uno recorra la mitad de la distancia que lo separa del otro.

Bibliografía[12]

AGUILAR, L.- FERNÁNDEZ, J. M.- GARRIDO, J. M.- LLISTERRI, J.- MACARRÓN, A.MONZÓN, L.- RODRÍGUEZ, M. A. (1994a) "Diseño de pruebas para la evaluación de habla sintetizada en español y su aplicación a un sistema de conversión de texto a habla", *Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, Córdoba, 20-22 de julio de 1994. http://liceu.uab.es/~joaquim/publicacions/cordoba_94.html

AGUILAR, L.- FERNÁNDEZ, J. M.- GARRIDO, J. M.- LLISTERRI, J.- MACARRÓN, A. - MONZÓN, L.- RODRÍGUEZ, M. A. (1994) "Evaluation of a Spanish text-to-speech system", *Conference Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis.* September 12-15 1994, Mohonk Mountain House, New Paltz, NY. pp. 207-210.
http://liceu.uab.es/~joaquim/publicacions/newyork_94.html

AGUILAR, L.- GARRIDO, J. M.- LLISTERRI, J. (1997) "Incorporación de conocimientos fonéticos a las tecnologías del habla", in SERRA, E.- GALLARDO, B.- VEYRAT, M.- JORQUES, D.ALCINA, A. (Eds.) *Panorama de la investigació lingüística a l'Estat Espanyol.* Actes del I Congrés de Lingüística General. Volum III. Comunicacions: Fonètica i Fonologia. Semàntica i Pragmàtica.
València: Universitat de València. pp. 5-13.
http://liceu.uab.es/~joaquim/publicacions/valencia_94.html

Base de datos oral del español Albayzín. Universitat Politècnica de València, Universidad Politécnica de Madrid, Universidad de Granada, Universitat Aut noma de Barcelona, Universitat Politècnica de Catalunya. 5 CD-ROM. 1999.

BLECUA, B.- ACÍN, V. (1995) "Propuesta de un modelo de intensidad vocálica del castellano y el catalán aplicable a un sistema de conversión de texto a habla", *Procesamiento del Lenguaje Natural, Revista* nº 17: 257-271.
http://www.sepln.org/revistaSEPLN/revista/17/17Pag257.pdf

CASACUBERTA, F.- GARCÍA, R.- LLISTERRI, J.- NADEU, C.- PARDO, J. M.- RUBIO, A. (1992) " Desarrollo de corpus para investigación en tecnologías del habla (Albayzín)", *Procesamiento del Lenguaje Natural*, *Boletín nº* 12: 35-42.
http://www.sepln.org/revistaSEPLN/revista/12/12-Pag35.pdf

[11] Esta diferente valoración del trabajo tiene, naturalmente, sus repercusiones económicas.

[12] La bibliografía recoge únicamente, de acuerdo con la intención del trabajo, las contribuciones del Grupo de Fonética de la UAB. El lector interesado en las tecnologías del habla puede encontrar referencias más generales en
http://liceu.uab.es/~joaquim/speech_technology/tecnol_parla/speech_tech_general/refs_gen_tecnol_parla.html

de la MOTA, C.- RIERA, M. (2000) "Segre i Aneto. Les regles del català central", *I Jornades del Centre de Referència en Enginyeria Lingüística* (CREL), Institut d'Estudis Catalans, Barcelona, 4 i 5
d'abril de 2000. http://liceu.uab.es/publicacions/SFI_UAB_Transcriptor.pdf

DYBKJAER, L.- BERMAN, S.- BERENSEN, N. O.- CARLETTA, J.- HEID, U.- LLISTERRI, J. (2001) *Requirements Specification for a Tool in Support of Annotation of Natural Interaction and Multimodal Data*. ISLE Natural Interactivity and Multimodality Working Group. D11.2. July 2001.
http://isle.nis.sdu.dk/reports/wp11/D11.2-ISLE-29.7.2001-F.pdf

ESTRUCH, M. (2000) "Évaluation de l'algorithme de stylisation mélodique MOMEL et du système de codage symbolique INTSINT avec un corpus de passages en catalan", *TIPA -Travaux Interdisciplinaires du laboratoire Parole et langage d'Aix-en-Provence* 19: 45-61.

GARRIDO, J. M. (1991) *Modelización de patrones melódicos del español para la síntesis y el reconocimiento*. Bellaterra: Departament de Filologia Espanyola, Universitat Aut noma de Barcelona. http://liceu.uab.es/juanma/Web/Postscript/Garrido91.ps

GARRIDO, J. M. (1996) *Modelling Spanish Intonation for Text-to-Speech Applications.* Ph.D. Thesis. Departament de Filologia Espanyola, Facultat de Lletres, Universitat Aut noma de Barcelona.
http://liceu.uab.es/juanma/tesis.html

GARRIDO, J. M. (2001) "La estructura de las curvas melódicas del español: propuesta de modelización", *Lingüística Española Actual* 23, 2: 173-209.

GARRIDO, J. M..- ORTÍN, I.- QUAZZA, S.- SALZA, P. L.- MANCINI, F. (2000) "Desarrollo de un módulo de asignación de parámetros prosódicos para la versión en español del sistema de conversión texto-habla ACTOR®", *Procesamiento del Lenguaje Natural*, *Revista nº* 26: 183-190.
http://www.sepln.org/revistaSEPLN/revista/26/garrido-alminana.pdf

LLISTERRI, J. (1997) "Experiències de col.laboració amb empreses en l'àmbit de les humanitats", *F rum de la Recerca: "Com fer convenis de col.laboració entre la UAB i les empreses. Què poden aprendre de l'experiència?"*, Universitat Aut noma de Barcelona, 22 d'octubre de 1997.
http://liceu.uab.es/~joaquim/publicacions/Forum_recerca.htm

LLISTERRI, J. (2001a) "El habla como medio de acceso a la Sociedad de la Información", *La Musa Digital* 1 (Monográfico: El impacto social de las nuevas tecnologías. La Sociedad de la Información).
http://www.uclm.es/ab/humanidades/lamusa/paginas/monografico/Llister ri.htm

LLISTERRI, J. (2001b) "La conversión de texto en habla", *Quark. Ciencia, Medicina, Comunicación y Cultura* 21: 79-89.
http://liceu.uab.es/~joaquim/publicacions/Quark2001/CTH_Quark_01.pdf

LLISTERRI, J.- AGUILAR, L.- GARRIDO, J. M.- MACHUCA, M. J.- MARÍN, R.- de la MOTA, C.- RÍOS, A. (1999) "Fonética y tecnologías del habla", in BLECUA, J.M.- CLAVERÍA, G.SÁNCHEZ, C.- TORRUELLA, J. (Eds.) *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona: Seminario de Filología e Informática, Departamento de Filología Española, Universidad Autónoma de Barcelona - Editorial Milenio. pp. 449-479.
http://liceu.uab.es/~joaquim/publicacions/Fonetica_TecnolHabla.pdf

LLISTERRI, J.- FERNÁNDEZ, N.- GUDAYOL, F.- POYATOS, J. J.- MARTÍ, J. (1993) "Testing user's acceptance of Ciber232, a text to speech system used by blind persons", in GRANSTRÖM, B.HUNNICUTT, S.- SPENS, K.-E. (Eds.) *Speech and Language Technology for Disabled Persons*. Proceedings of an ESCA Workshop. Sotckholm, Sweden, May 31-June 2, 1993. pp.203-206. http://liceu.uab.es/~joaquim/publicacions/Stockholm_93/stockholm_93. html

LLISTERRI, J.- GARRIDO, J. M. - RÍOS, A.- JIMÉNEZ, E. (2000) "Models pros dics per a la conversió de text a parla", *I Jornades del Centre de Referència en Enginyeria Lingüística* (CREL), Institut d'Estudis Catalans, Barcelona, 4 i 5 d'abril de 2000. http://liceu.uab.es/publicacions/SFI_UAB_Models_prosodics.pdf

MACHUCA, M. J.- BUENO, L.- CALONGE, R.- ESTRUCH, M.- RIERA, M. (2000a) "Corpus de diàleg", *I Jornades del Centre de Referència en Enginyeria Lingüística* (CREL), Institut d'Estudis Catalans, Barcelona, 4 i 5 d'abril de 2000. http://liceu.uab.es/publicacions/SFI_UAB_Corpus_Dialeg.pdf

MACHUCA, M. J.- BUENO, L.- CALONGE, R.- ESTRUCH, M.- RIERA, M. (2000b) "Eines de reconeixement i prototip de conversa oral", *I Jornades del Centre de Referència en Enginyeria Lingüística* (CREL), Institut d'Estudis Catalans, Barcelona, 4 i 5 d'abril de 2000. http://liceu.uab.es/publicacions/SFI_UAB_Disseny_prototip.pdf

MARÍN, R. (1994) "Diseño y evaluación de un modelo de duración vocálica del español para la síntesis del habla", *Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, Córdoba, 20-22 de julio de 1994.

MENGEL, A. - DYBKJAER, L., GARRIDO, J.M. - HEID, U.- KLEIN, M. - PIRRELLI V. POESIO, M. - QUAZZA, S. - SCHIFFRIN, A. - SORIA, C. (2000) *MATE Dialogue Annotation Guidelines*. MATE Deliverable D2.1. 8 January 2000. http://www.ims.unistuttgart.de/projekte/mate/mdag/

MORENO, A.- POCH, D.- BONAFONTE, A.- LLEIDA, E.- LLISTERRI, J.- MARIÑO, J. B.- NADEU, C. (1993) "ALBAYZIN Speech Database: Design of the Phonetic Corpus" in *Eurospeech'93. 3rd European Conference on Speech Communication and Technology.* Berlin, Germany, 21-23 September 1993. Vol. 1 pp. 175-178.

PACHÈS, P.- DE LA MOTA, C.- RIERA, M.- PEREA, P.- FEBRER, A.- ESTRUCH, M.GARRIDO, J. M.- MACHUCA, M. J.- RÍOS, A.- LLISTERRI, J.- ESQUERRA, I.- HERNANDO, J.-
PADRELL, J.- NADEU, C. (2000) "SEGRE: An Automatic Tool for Grapheme-to.Allophone Transcription in Catalan", in Ó CRÓINÍN. D. (Ed.) *Proceedings of the Workshop on Developing Language Resources for Minority Languages: Reusability and Strategic Priorities (LREC-2000 Second International Conference on Language Resources and Evaluation)*. Athens, 30 May 2000. pp.
52-61. http://liceu.uab.es/~joaquim/publicacions/Paches_et_al_2000.pdf

PUIGVÍ, D.- JIMÉNEZ, D.- FERNÁNDEZ, J. M. (1994) "Parametrización de las pausas ortográficas en castellano. Aplicación a un conversor de texto a habla", *Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, Córdoba*, 20-22 de julio de 1994.

RIERA, M. (1999) "Definició d'una unitat d'àmbit local per a la generació automàtica de l'entonació en català", *Actes del I Congrés de Fonètica Experimental*. Tarragona, 22, 23 i 24 de febrer de 1999. Universitat Rovira i Virgili - Universitat de Barcelona. pp. 295-301.

RIERA, M. (2001) *Anàlisi acústica dels moviments tonals del grup accentual en català*. Treball d'investigació de Tercer Cicle - Programa de doctorat de Lingüística: tractament informàtic del llenguatge. Departament de Filologia Espanyola, Universitat Aut noma de Barcelona. http://liceu.uab.es/~montse/pubs/Riera2001.pdf

RIERA, M.- JIMÉNEZ, E. (2000) "Corpus pros dic", *I Jornades del Centre de Referència en Enginyeria Lingüística* (CREL), Institut d'Estudis Catalans, Barcelona, 4 i 5 d'abril de 2000. http://liceu.uab.es/publicacions/SFI_UAB_Corpus_Prosodic.pdf

RÍOS, A. (1993) "La información lingüística en la transcripción fonética automática del español", *Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural* 13: 381-387. http://www.sepln.org/revistaSEPLN/revista/13/13-Pag381.pdf

RÍOS, A. (1999) La transcripción fonética automática del diccionario electrónico de formas simples flexivas del español: Estudio fonológico en el léxico. *Estudios de Lingüística Española* 4. http://elies.rediris.es/elies4/

WEGENER, R.- MARTIN, J. C.- DYBKJAER, L.- MACHUCA, M. J.- BERNSEN, N.O.CARLETTA, J.- HEID, U.- KITA, S.- LLISTERRI, J.- PELACHAUD, C.- POGGI, I.REITHINGER, N.- van ELSWIJKS, G.- WITTENBURG, P. (2002) *Survey of Multimodal Coding Schemes and Best Practice*. ISLE Natural Interactivity and Multimodality. Working Group Deliverable D9.1. February 2002. http://isle.nis.sdu.dk/reports/wp9/D9.17.3.2002-F.pdf