PROYECTOS EN CURSO

TANGORA/E, un reconocedor del habla para el castellano

J.González, J.Macías, M.A.Palma, F.Palou, M.Tros de Ilarduya

Centro de Tecnología de la Lengua
Edificio S-3, EXP0'92
41092 Sevilla

Introducción

El Centro de Tecnologia de la Lengua está desarrollando para el español el reconocedor de habla Tangora/E. Este desarrollo es parte de una tarea más amplia que IBM lleva a cabo en Estados Unidos y en varios países europeos (Italia, Alemania, Francia y el Reino Unido), en distinto grado de avance, dependiendo de cuándo comenzaran y de los recursos dedicados. Por ejemplo, se dispone ya de sistemas de 20.000 palabras para el inglés americano, el italiano y el francés. El desarrollo de Tangora se inició, a principios de los años 70, en el T. J. Watson Research Center, de Nueva York.

En el Centro de Tecnología de la Lengua (CTL) de Sevilla comenzamos a trabajar en el proyecto a mediados de 1990. A finales de ese año habíamos conseguido un sistema que reconocía 50 palabras y, en junio de 1991, otro de 1.200 palabras. En mayo de 1992 se comenzó la evaluación de un sistema de 6.000 palabras.

Nuestra meta es llegar a vocabularios de 20.000 palabras, como se está haciendo en otras lenguas.

Descripción

El sistema Tangora español de 1.200 palabras funciona hoy en un PC AT de IBM, dotado de un micrófono, un conversor analógico-digital, un amplificador y altavoz y tarjetas procesadoras construidas especialmente para el reconocedor. El sistema de 6.000 palabras funciona en un ordenador RS-6000 de IBM con una sola tarjeta procesadora especial más otra de digitización estándar.

De momento y hasta que se construyan, a coste razonable, procesadores capaces de resolver, en tiempo real, el cálculo necesario para el habla continua, unas cien veces más potentes que los actuales, es necesario hablar con cortas pausas entre palabras, que ahorran la tarea de delimitación, aunque los algoritmos vienen a ser los mismos que para el habla continua. Bastan unos pocos centisegundos para producir esta pausa. La pausa no elimina por completo el proceso de coarticulación y se puede mantener la entonación normal.

En la historia del desarrollo de Tangora en Estados Unidos se han logrado excelentes resultados con habla continua: en 1976, prácticamente sin error, con un vocabulario de 250 palabras; en 1979, 9% de error en 1.000 palabras. Actualmente, la tasa de error entre palabras pronunciadas y palabras escritas esta situada entre el 2 y el 5% en los sistemas de 20.000 palabras.

Tangora tiene dos modos de uso: uno de pronunciación de palabra entera y otro de deletreo, en el que puede elegirse la pronunciación de la letra o de una palabra asociada a ella previamente. En modo "deletreo", podría dictarse "rosa" como: Roma, Oviedo, Sevilla, Alemania, utilizando las convenciones telefónicas más usuales o el alfabeto de la Unión Internacional de Telecomunicaciones: Romeo, Oscar, Sierra, Alfa.

El sistema necesita entrenamiento individual, aunque hemos observado que funciona bien sin entrenamiento cuando el número de palabras de su diccionario es pequeño. El entrenamiento se basa en un texto de cien frases de diez palabras, estudiadas para que se pronuncien todos los pseudofonemas, que llamamos fonos, un número suflciente de veces y en distintos contextos. Por término medio, el tiempo de entrenamiento es de veinte minutos para cada locutor. Hay 500 palabras distintas en el texto de entrenamiento.

El vocabulario de 1.200 palabras del español, sobre un texto nuevo especializado, cubre algo más del 76% de las palabras de dicho texto, entendiendo como "palabra" cada forma flexionada, no cada lema. El de 6.000 palabras cubre alrededor del 90%. El corpus sobre el que. hemos trabajado es de unos seis millones de palabras, de las cuales son distintas 78.000. Hemos procesado parte de la sección de Economía y editoriales del diario EL PAIS.

Los datos de cobertura de nuestro vocabulario sobre un texto nuevo especializado son:

- 100 palabras cubren el 56%

- 1.000 palabras cubren el 76%

- 5.000 palabras cubren el 90%
- 10.000 palabras cubren el 94%
- 15.000 palabras cubren el 96%
- 20.000 palabras cubren el 97%

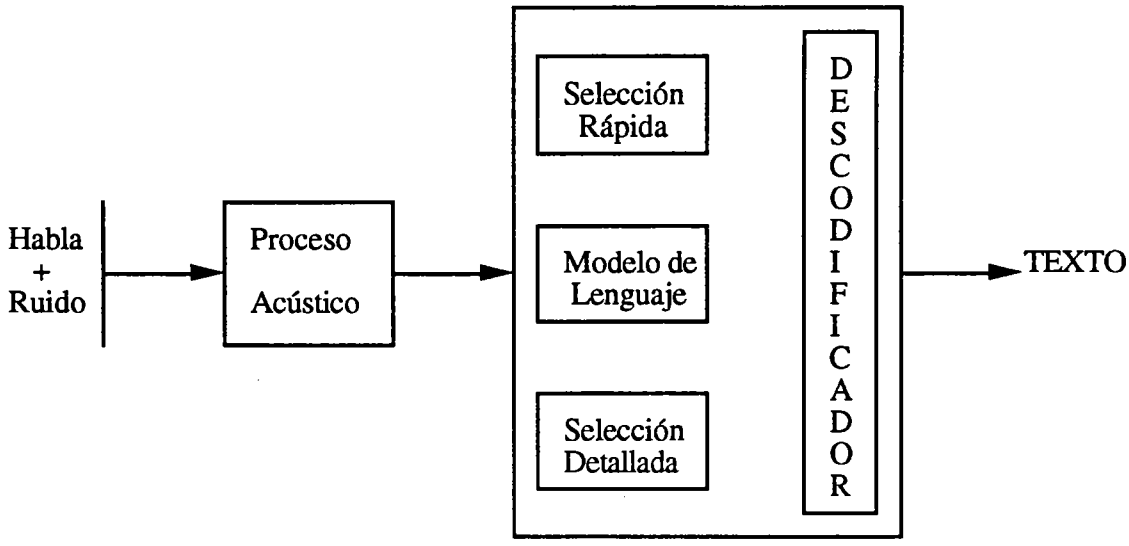
Organización del Sistema

El reconocedor de habla Tangora está basado en técnicas relacionadas con la teoría de información y la teoría de detección óptima. En estas técnicas se considera una "fuente" que emite "símbolos", un "canal" por el que se transmiten y que los hace susceptibles de distorsión y de contaminación por ruido y un "receptor" que capta esta información deteriorada y, a partir de la misma, infiere los "símbolos" que la "fuenten" ha generado.

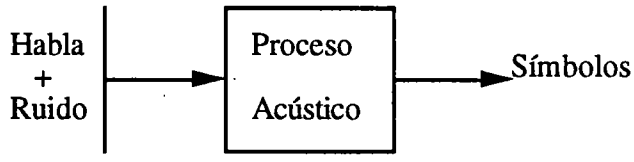
El "receptor" decide qué simbolos han sido transmitidos realmente, haciendo óptima su probabilidad de ocurrencia, a partir de la información captada y del conocimiento de la probabilidad de generación de los "símbolos" por la "fuente" y de las características del "canal".

En nuestro reconocedor, la "fuente" es el locutor y sus "símbolos", las diferentes palabras que constituyen su discurso. El "canal" modela la acústica de la sala, las características del micrófono, el ruido ambiental y el de la electrónica asociada, la distorsión producida en el proceso acústico, ya por errores en el cálculo, ya por la cuantización interna de la información, etc.

El "receptor" es nuestro descodificador propiamente dicho. Trata de decidir lo que el locutor ha pronunciado a partir de esa información proporcionada por el "canal". Como decimos antes, se trata de hacer óptima la probabilidad de que las palabras reconocidas sean realmente las pronunciadas por el locutor y esto se hace considerando no sólo cada palabra de forma aislada sino, en principio, la frase entera, desde la activación a la desactivación del micrófono.



Nuestro descodificador recibe información acústica del procesador acústico, informacion proporcionada por los procesos que denominamos "Selección Rápida" y "Selección Detallada" sobre la probabilidad de que cada palabra del vocabulario haya sido pronunciada, dada la información acústica recibida y, por último, una estimación de la probabilidad de ocurrencia en la lengua de una secuencia de palabras dada. Con todo ello, el descodificador estima la secuencia de palabras más probable pronunciada por el locutor. Es, por tanto, un proceso probabilístico.

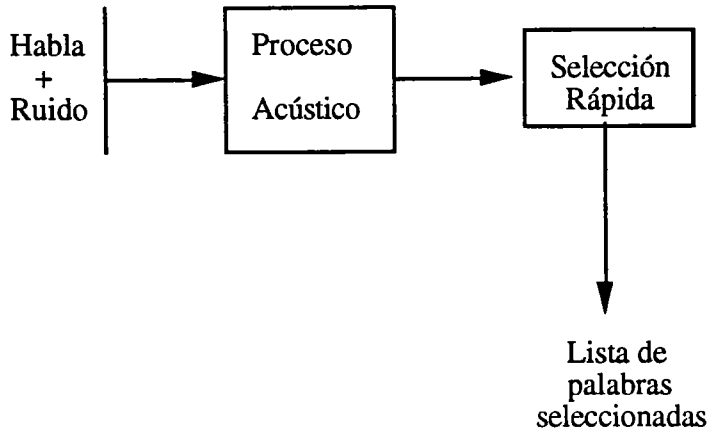


La señal captada por el micrófono es filtrada para eliminar frecuencias superiores a 8 khz., muestreada a 20.000 muestras por segundo y cuantizada a 12 bits. El espectro instantáneo de la señal es calculado en 20 bandas espectrales cien veces por segundo y, mediante un proceso denominado "cuantización vectorial", se compara con otros obtenidos durante el entrenamiento, clasificándose en uno de 200 posibles espectros.

Combinado con este proceso, tiene lugar uno de normalización frente a las condiciones cambiantes del ruido, del nivel de la señal y de las características acústicas de la habitación y del micrófono.

El resultado es una sucesión de valores numéricos a un ritmo de 100 por segundo, que representan la información acústica correspondiente a lo pronunciado por el locutor.

Selección Rápida



En este proceso se lleva a cabo una comparación aproximada con todas las palabras del diccionario contenido en el sistema. La Selección Rápida produce una corta lista de palabras candidatas. Para calcular la probabilidad de que un modelo acústico corresponda a una palabra, se ha modelado previamente la palabra como una secuencia de fonos, que tiene la apariencia de una transcripción fonética. Los fonos están representados por máquinas de estados finitos (modelos ocultos de Markov). Estos fonos corresponden a los fonemas y alófonos de la lengua. Se han definido 51 fonos para el español. La palabra es, pues, una sucesión de fonos.

Existen homógrafos que pueden tener más de una representación, caso poco frecuente en español, por ejemplo "y", conjunción e "y griega". Por el contrario, hay frecuencia de homófonos, con distinta grafía "ha", "a"; "d"(letra), "de"; "el", "él", que solamente puede desambiguar el "Modelo de LengUaje".

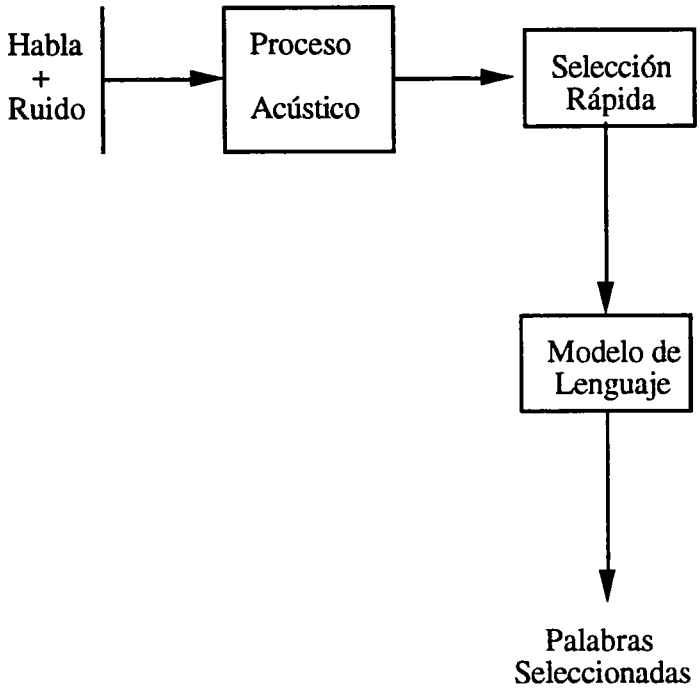
La ristra de fonos correspondiente a una pronunciación de una palabra dada se llama forma-base.

Modelo de Lenguaje

Asigna probabilidades a secuencias de palabras, independientemente del proceso acústico. El "Modelo de Lenguaje" es un modelo de Markov en donde la probabilidad de una palabra viene condicionada por las palabras precedentes.

Se ha preferido el modelo estadístico a la constricción de una gramática, aunque se está tomando en cuenta la posible "ayuda" de gramáticas simples al modelo estadístico:

P(se llegó a un acuerdo ...) =
P(se) x P(llegó | se) x P(a | se llegó) x P(un | se llegó a) x ...



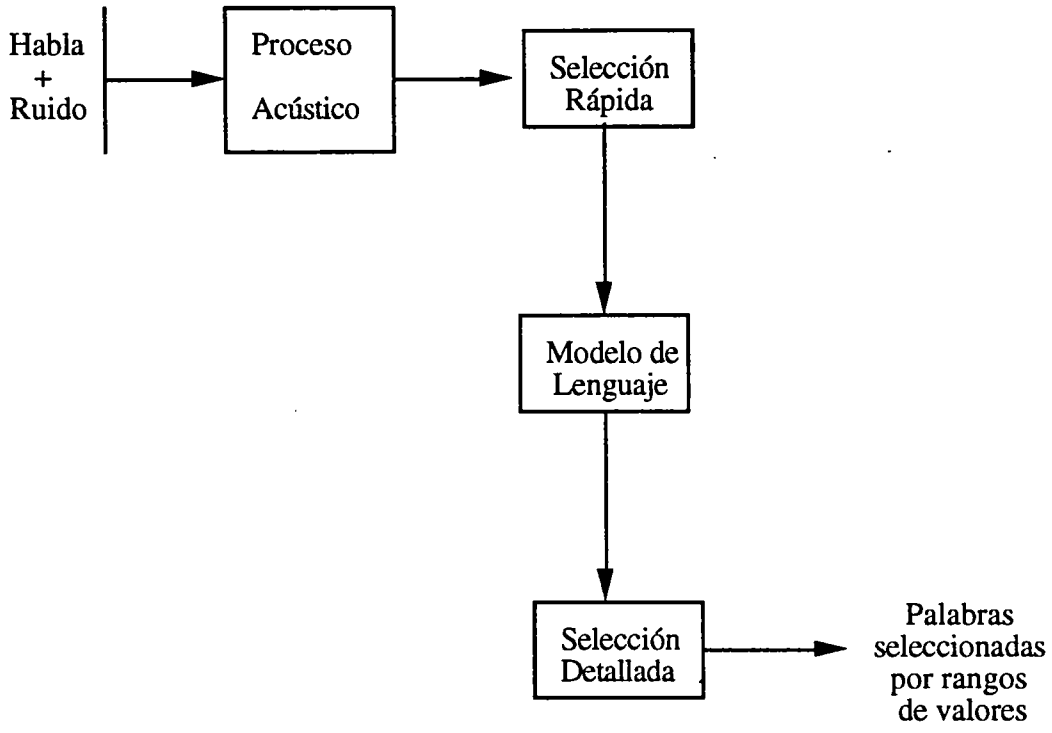
En el modelo de "trigramas" se trunca la historia anterior a dos palabras (modelo de Markov de segundo orden). Las probabilidades se derivan de la cuenta de frecuencia relativa sobre un texto representativo. Las probabilidades son suavizadas para eliminar los ceros, esto es, la inexistencia de esa secuencia en el corpus. Se permite cualquier secuencia de palabras. En la versión de 1.200 palabras hemos tomado como texto el corpus mencionado al principio. En la de 6.000 palabras, en la que estamos trabajando, las fuentes de texto serán varias con lo que esperamos que el modelo sea más rico.

El "Modelo de Lenguaje" reduce la lista de palabras obtenidas por la "Selección Rápida", asignándoles valores normalizados de probabilidad.

Selección Detallada

Este proceso consiste en hallar una equivalencia más exacta sobre un pequeño grupo de palabras seleccionadas. Es más caro, hablando en terminos de cómputo, que la "Selección" rápida. Se basa

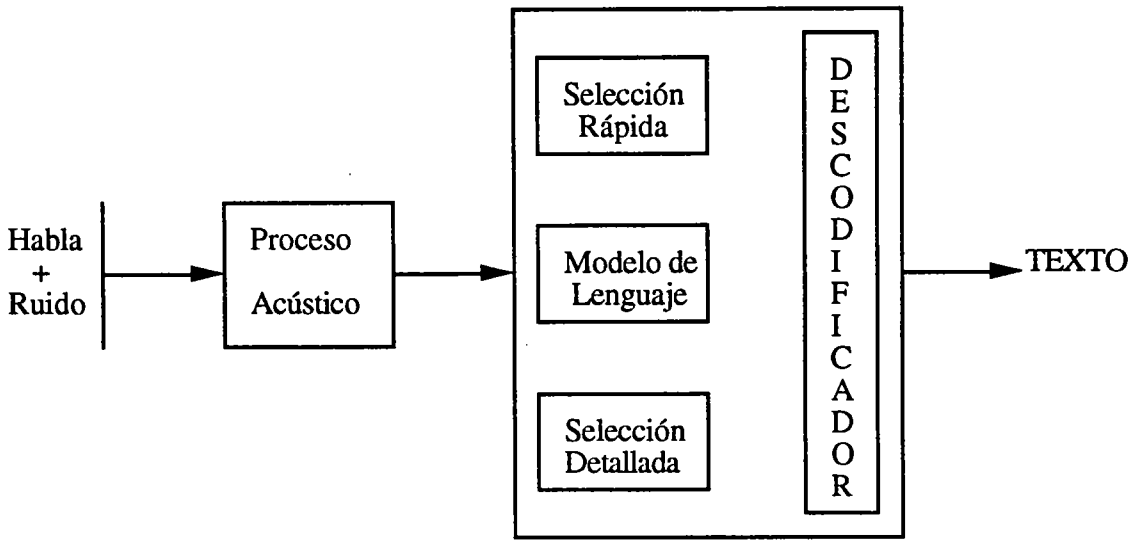
en modelos acústicos fenónicos, término que significa una resolución más alta que la de los modelos fonéticos usados en la "Selección Rápida". Estos modelos han sido generados a partir de pronunciaciones de las mismas palabras emitidas por varios locutores. Son modelos de palabras enteras, no basados en transcripciones fonéticas convencionales, sino en la pronunciación real de cada palabra. Por ello son más precisos que los modelos utilizados en la "Selección Rapida".



Para el español, hemos utilizado diez voces masculinas, con pronunciación uniforme, pero con distintos tonos. Nuestros locutores han leído el texto de entrenamiento y las palabras del vocabulario.

El modelo de cada palabra está constituido por una sucesión de modelos básicos escogidos de un repertorio de 200. Puesto que el repertorio no es muy grande, el número de parámetros a estimar es relativamente pequeño e independiente del tamaño del vocabulario. Este aspecto es crucial al facilitar la labor de entrenamiento de los modelos acústicos por el locutor.

Descodificador



El descodificador controla el funcionamiento del reconocedor. En síntesis, el descodificador funciona como sigue:

La información acústica proporcionada por el procesador acústico es utilizada para obtener una lista de las palabras más probables pronunciadas en el proceso de "Selección Rápida".

El "Modelo de Lenguaje" se tiene en cuenta a la hora de asociar una puntuación a cada palabra. El objetivo de esta fase es limitar el número de palabras sobre las que se va a realizar el proceso de "Selección Detallada" de la siguiente fase. Los parámetros del proceso se ajustan para que la secuencia correcta de palabras no quede descartada aunque no tenga la mayor probabilidad.

En la fase siguiente se analizan las palabras candidatas en cada secuencia mediante el proceso de "Selección Detallada". La puntuación dada a cada palabra es función no solo del proceso de "Selección Detallada" sino también del "Modelo de Lenguaje" y de la puntuación dada durante la "Selección Rápida". Se exploran posibles alternativas de secuencias de palabras y se da como buena aquella con mayor puntuación acumulada. En la práctica no se analizan todas las secuencias posibles de palabras, ya que tendría un coste prohibitivo. Aquellas secuencias con una puntuación acumulada por debajo de un umbral, se descartan.

Conforme avanza el proceso de descodificación, se van fijando sucesivamente las palabras del discurso, sin esperar a que éste termine. La exploración continúa a partir de las palabras consideradas como firmes.

En la práctica, este proceso de verificación de hipótesis de secuencias de palabras, queda limitado típicamente a los cinco últimos segundos de discurso.

Resultados

Los resultados obtenidos son altamente esperanzadores. Sobre un texto de 508 palabras, tomado al azar de la sección de Economía del diario EL PAIS y con la única particularidad de estar completamente cubierto por el vocabulario, la tasa de error se situó entre 2,9 y 5,5% para los cinco locutores que se consideraron en las pruebas de la version de 1.200 palabras. En la versión de 6.000 palabras se consideraron 8 locutores de los que dos de ellos usaban el sistema por primera vez. La tasa de error se situó entre 2,95 y 18,50% (media 8,29%) cuando se consideraron los 8 locutores, y entre 2,95 y 11,81% (media 5,22%) cuando se excluyeron los locutores sin experiencia. En todos los casos las pruebas se realizaron con un modelo de lenguaje ficticio que aparece como neutro en el proceso de descodificación. Por esta razón no consideramos como errores la sustitución de palabras por sus homófonos.

En la versión de 1.200 palabras introdujimos un modelo de lenguaje obtenido a partir de un corpus de aproximadamente 1,5 millones de palabras de la sección de Economía de EL PAIS que no incluye el texto de prueba. La tasa de error se situó entre 1,4 y 3,2% para los mismos locutores empleados para las pruebas sin modelo de lenguaje pero incluyendo esta vez como error las sustituciones por homófonos. En la versión de 6.000 palabras hemos realizado un modelo provisional utilizando un corpus de 5 millones de palabras obtenidos de textos del periódico Heraldo de Aragon. La tasa de error se sitúa entre 0,39 y 7,28% (media 2,87%) para todos los locutores y entre 0,39 y 3,15% (media 1,14%) para los locutores con experiencia. Esperamos que se mejoren estas cifras con un modelo de lenguaje más elaborado.

Futuras investigaciones

Actualmente (mayo 1992) estamos mejorando los modelos acústicos y preparando un modelo de lenguaje para la versión de 6.000 palabras.

Puesto que el modelo de lenguaje se basa en "trigramas", el tamaño del corpus utilizado para estimar el modelo debería crecer, en una primera aproximación, con el cubo del tamaño del vocabulario. En la actualidad disponemos de un corpus de unos 80 millones de palabras procedente

de diversas fuentes que suponemos va a ser más que suficiente para la versión de 6.000 palabras. No estamos tan seguros de que lo sea para la versión futura de 20.000 palabras por lo que continuamos adquiriendo y depurando texto para incorporarlo a nuestro corpus.

Queremos explorar la posibilidad de sintetizar modelos acústicos fenónicos de palabras a partir de modelos fenónicos de alófonos. Esto nos permitiría la generación de modelos de palabras nuevas sin necesidad de que éstas tengan que ser pronunciadas. Pensamos que la información acústica contenida en las grabaciones de que disponemos nos va a permitir establecer un inventario de alófonos adecuado a esta síntesis.

Por último, queremos estudiar los problemas asociados a la ergonomía de un reconocedor de voz. Puesto que nuestra motivación a más largo plazo es la sustitución del teclado del terminal del ordenador, la definición de la forma de comunicación con el usuario y el conocimiento del rango de aplicaciones de validez de la tecnologla son de capital importancia. Por muy perfecto que sea el reconocedor, van a existir errores de reconocimiento, tal y como ocurre con un interlocutor humano. La detección de estos errores (la palabra reconocida incorrectamente es ortográficamente correcta) y su corrección (¿sin teclado?), van a presentar un desafío a nuestra imaginación y a la de otros equipos investigadores. Con objeto de investigar estos problemas y situarnos dentro de la realidad y no simplemente en el ambiente ideal del laboratorio, nuestro grupo se propone aplicar la utilización del reconocedor a un problema práctico, en colaboración con algún grupo elegido de usuarios potenciales. Estamos considerando el dictado de informes radiológicos en hospitales como primer caso práctico de estudio.

Agradecimiento

Agradecemos a los periódicos El País y Heraldo de Aragón, así como a las editoriales Ediciones Doyma S.A, Editorial Labor S.A., Editorial Planeta S.A. y Tusquets Editores S.A., su desinteresada colaboración al proporcionarnos parte del material empleado en la confección del corpus textual utilizado en estas investigaciones.

Bibliografía

A. AVERBUCH et al., "Experiments with the Tangora 20,000 word speech recognizer", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, (Dallas, Texas), pags. 701-704, Abril 1987.

P. ALTO et al., "Experimenting Natural-Language Dictation with a 20,000 word speech recognizer", IEEE CompEuro 89, Hamburgo, pags. 2/78-2-81, Mayo 1989

L.R. BAHL et al., "Acoustic Markov models used in the Tangora speech recognition system", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pags. 497-500,1988.