Diseño de un base de datos tipo SpeechDat para el idioma gallego *

Begoña González Rei
Departamento de Filoloxía Galega, Universidade de Santiago de Compostela.
fgbego@usc.es

Carmen García Mateo
E.T.S. E. Telecomunicación, Universidade de Vigo
carmen@gts.tsc.uvigo.es

Resumen En la actualidad se está recogiendo una base de datos oral telefónica para el gallego. Se inserta dentro de las iniciativas de recogida de bancos de datos orales estándar del SpeechDat, concretamente en el proyecto SpeechDat(M), que recoge muestras de voz de aproximadamente 1000 hablantes que llaman desde un teléfono fijo (FDB, Fixed Data Base). El diseño de SpeechDat pretende ser general, es decir, por un lado válido para todas las lenguas representadas y, por otro, permitir distintas aproximaciones al problema del reconocimiento del habla (reconocimiento de palabras aisladas, conectadas y del discurso continuo). En este trabajo se describe al diseño de esta base de datos SpeechDat, reflejando la problemática de su adaptación al gallego y las novedades incluidas. Los temas que se van a abordar en esta comunicación son: a) el diseño de la base de datos, teniendo en cuenta los requisitos genéricos para todos los idiomas y la propia idiosincrasia del gallego; b) la adquisición de las muestras orales telefónicas, con una descripción tanto de la metodología como del sistema de grabación; c) el etiquetado del material oral, con la descripción del nivel de representación de los datos.

1 *Introducción*

Los sistemas de reconocimiento de habla considerados como "estado del arte"en la comunidad científica se basan en un aprendizaje a partir de datos. Para su desarrollo es necesario disponer de grandes bases de datos de voz que reflejen en buena medida las condiciones de funcionamiento de la aplicación diseñada.

* Este trabajo ha sido parcialmente por la CICYT con del proyecto FEDER 1FD97-0077 "Sistemas de diálogo para el acceso telefónico a servicios telemáticos "

En Europa se están realizando grandes esfuerzos para generar recursos lingüísticos en las lenguas mayoritarias de la comunidad Europea y parte de las lenguas minoritarias. Proyectos como ESPRIT SAM, SpeechDat [4], SpeechDat II [5], Vodis, SpeechDat Car [3], SALA [1] o últimamente SpeechCom, han tenido o tienen como objetivo la generación de tales recursos. A lo largo de los años el proceso de diseño y captura se ha ido refinando llegándose a unas especificaciones estándar que hacen que con decir que la base de datos es tipo *SpeechDat* se sepa de qué se está hablando.

Las bases de datos SpeechDat son del tipo telefónico orientadas al desarrollo de aplicaciones de teleservicios. En estos momentos existen más de 14 de estas bases de datos para la red telefónica fija. El idioma gallego no dispone actualmente de una de estas bases de datos que como decimos es condición casi indispensable para desarrollar una aplicación de reconocimiento de habla a través de la línea telefónica. En este artículo se describe el diseño del material que compondrá la base de datos SpeechDat en el idioma gallego. Este diseño se ha hecho siguiendo escrupulosamente los documentos del proyecto [6, 2]. Asimimismo se ha completado con material adicional sobre todo en lo concerniente a la recogida de *números* y *frases*.

2 *DESCRIPCIÓN DE LOS CONTENIDOS DE LA BASE DE DATOS.*

2.1 Material obligatorio de una base de datos tipo SpeechDat.

El objetivo principal de la base de datos es ser válida y práctica para su aplicación en todo tipo de teleservicios. En consecuencia, el vocabulario de la base de datos contiene una gran variedad de material, entre el que destaca:

1. Frases fonéticamente ricas, que procuran cubrir todas las realizaciones fonéticas de la lengua. En total se reservan 10 ítems del cuestionario a estas frases. Se trata de frases entre 8 y 12 palabras mayores de 3 caracteres, extraídas de un gran corpus textual de estilo periodístico.

2. Palabras aisladas fonéticamente ricas. hay cuatro ítems reservados a este tipo de material. Se trata de palabras reales y no de logátomos, y su tamaño oscila entre los 10 y los 15 caracteres.

3. Deletreo de palabras y cadenas de letras. En total la base de datos incluye tres ítems con este material, dos leídos (uno con el deletreo de una población, y otro de cadenas de letras en las que se potencian aquellas menos comunes) y uno espontáneo.

4. Material orientado a las aplicaciones del reconocedor. Dentro de este tipo se incluyen cuatro ítems con los comandos generales de los sistemas automáticos (tres de comandos aislados y uno con el comando en 'word spotting'). También se pueden considerar dentro de este grupo a los ítems consistentes en nombres propios, con el fin de recoger posible material de los directorios de estos sistemas. Así, tenemos cuatro puntos reservados a nombres propios de persona, dos leídos (un nombre de pila, un apellido) y dos espontáneos (un nombre de pila, un nombre completo). Dos puntos se corresponden con nombres de lugar (uno leído y uno espontáneo), y un ítem contiene el nombre de una empresa o institución.

5. Números. Se incluyen en este grupo varios ítems reservados a dígitos aislados, a cadenas de dígitos y a cualquier número natural. Otro subgrupo lo forman otros números expresados en distintos formatos. La base de datos consta de dos números de teléfono, una tarjeta de crédito leída, un DNI espontáneo, dos códigos (uno el de identificación de la planilla) y una cantidad de dinero. Se trata de expresiones fundamentales en la mayoría de los telesistemas automáticos: transacciones bancarias, compras por teléfono, etc.

6. Expresiones de la fecha y de la hora, fundamentales en muchos sistemas automáticos, como por ejemplo de consulta de horarios o de reserva telefónica de billetes. En total hay tres ítems reservados a fechas, uno espontáneo y dos leídos, y entre ellos uno con expresiones relativas o generales de la fecha. En lo relativo a las expresiones de la hora, un elemento consiste en la lectura de una hora presentada en formato analógico y no digital, y el otro en una fecha espontánea.

El material oral recogido en esta base de datos es en parte leído y en parte espontáneo. En total se diseñaron un número total de 47 realizaciones por hablante y llamada, incluyendo realizaciones leídas (36) y espontáneas (10). Los textos escritos que se presentan al hablante para que los lea se recogen en un cuestionario con 35 puntos al que hay que añadir el código de cada una de estas planillas. A parte, se contemplan tanto en las hojas informativas como en el diálogo del sistema de grabación una serie de preguntas que conllevan respuestas espontáneas aunque guiadas. El diseño de estas preguntas fue muy cuidadoso con la finalidad de asegurar la máxima naturalidad en las respuestas, ya que aunque este objetivo preside el diseño de toda la base de datos, en cierta medida es previsible una pronuncia formal y una articulación más cuidada de los elementos presentados en forma escrita. Así, varios de los ítems diseñados recogen respuestas de los usuarios cuando piensan que el cuestionario ya ha acabado, pues se les indica que las preguntas correspondientes se deben únicamente a la necesidad de identificarlos para su participación en el sorteo.

El objetivo de obtener realizaciones lo más próximas posible a la realidad preside el diseño de toda la base de datos, y en consecuencia se tuvo especial cuidado en la selección y en la presentación de todos los elementos escritos: se buscaron frases claras y con expresiones habituales, extrayéndolas a este propósito de un corpus periodístico y no literario; el corpus de palabras fonéticamente ricas está constituido por palabras reales, existentes en la lengua, y no por logátomos, palabras inexistentes pero fonéticamente posibles; entre los números de teléfono los hay característicos de la red fija y propios de los móviles, con prefijos provinciales y 600 existentes en la realidad, y en el cuestionario se presentaron en distintos formatos (separados de dos en dos dígitos y de tres en tres) en previsión de que la forma escrita pudiese condicionar su lectura.

2.2 Material adicional propio de la SpeechDat en gallego.

En el diseño de la base de datos se respetaron escrupulosamente las especificaciones SpeechDat, incluso cuando los requisitos exigidos non concuerdan con la realidad del gallego. Así, por ejemplo, hay un total de tres ítems reservados al deletreo, aunque se trata de un material poco relevante para el reconocimiento en gallego, idioma en el que no se acude al deletreo con frecuencia, ni siquiera para desambiguar contextos, y su utilidad se restringe al reconocimiento de siglas. No obstante, se cuidó que el conjunto de realizaciones de las letras fuese suficiente y compensado, potenciándose en uno de los ítems las letras menos frecuentes.

A parte del material obligatorio se incluyó también material nuevo exclusivo de la base de datos SpeechDat gallega. Parte de este material adicional tiene su razón de ser en la orientación del proyecto al que se va a aplicar en primera instancia el reconocedor de voz entrenado con esta base de datos, pero la mayor parte de las novedades se deben a la propia realidad lingüística y sociolingüística del gallego. En total hay tres ítems leídos nuevos no contemplados en la tabla general de requisitos:

1. una frase fonéticamente rica (ítem S0)
2. un número natural (ítem N2) más de los obligatorios, ya que las frases y los números van a constituir el material fundamental para el entrenamiento del reconocedor, y
3. un apellido (ítem O8) como apoyo extra dentro del material propio de directorio.

Se incluyen también tres nuevos ítems espontáneos. Se trata de:

1. el número de un documento de identidad (ítem C6), formato no incluido entre los ítems destinados a la lectura, pero fundamental para la identificación de usuarios para, por ejemplo, transacciones bancarias telefónicas, consulta de saldo, etc.;
2. un nombre propio completo (ítem O9), aumentándose así el número de elementos del directorio de antropónimos, con la ventaja de una pronunciación más natural y fluida al tratarse en este caso de un nombre conocido (el de uno mismo);
3. un número de teléfono adicional (ítem C5).

El número de teléfono extra está motivado por la gran importancia que van a tener en la mayor parte de los teleservicios automáticos, pero la razón de incluirlo como elemento espontáneo está en la preocupación por recoger las posibles variaciones de pronunciación debidas a la diferencia entre lectura y producción espontánea y a los distintos grados de familiaridad del número. La mayor parte de los números de teléfono utilizados en sistemas automáticos serán números conocidos, que el hablante sabe de memoria y que, por lo tanto, reproduce con soltura; en consecuencia, aumenta la velocidad de producción y por lo tanto la coarticulación.

A parte de los ítems adicionales, se incluyeron novedades en el diseño de cada ítem, en los que se va muchas veces más allá de los simples requisitos del SpeechDat. Así, por ejemplo, entre las palabras y frases llave, a los 25 comandos obligatorios se añadieron 5 comandos nuevos: entre ellos hay varios que previenen nuevas posibilidades de aplicaciones generales en sistemas automáticos. Por un lado, "idioma"(aparte del comando obligatorio "galego", que se corresponde con la lengua de la base de datos), pensando en aplicaciones multilingua; por otra, "opcións", pensando en la posibilidad de programación de preferencias de usuario, y, finalmente, "axenda", pensando en la necesidad de este elemento en aplicaciones como la de consulta telefónica del correo electrónico. Otros dos comandos extra se deben a la idiosincrasia del gallego: uno de ellos, "inicio", se incluye por la necesidad de proporcionar un comando equivalente a "último"para acceder al primer elemento de cualquier directorio o menú; el otro se debe a las características particulares del gallego, en el que los adjetivos presentan alternancia de género (frente al inglés) y, por lo tanto, en muchas frases con comandos, en lugar de "último"puede ser más adecuada una variante "última"("última mensaxe", "última función", "última opción").

3 *DISEÑO EQUILIBRADO DE LA BASE DE DATOS.*

3.1 Material para el entrenamiento de reconocedores fonéticos.

La base de datos está destinada tanto al entrenamiento como a la validación del reconocedor. El conjunto total, por lo tanto, se divide en dos subcorpus, uno de "train"y otro de "test". La distribución de estos subcorpus para una BD de 1000 hablantes es de 800 para el entrenamiento y 200 para la validación.

El material recogido es común a los dos subcorpus, excepto en el caso de las frases fonéticamente ricas. En este caso se seleccionaron elementos distintos para evaluar el reconocimiento de realizaciones no entrenadas directamente, y se reservó un pequeño grupo de 200 frases comunes a entrenamiento y validación, para comparar la calidad del reconocimiento en los dos casos y determinar si la tasa de error en cada uno de ellos es aproximadamente la misma.

El conjunto del material oral recogido tiene que dar lugar a una base de datos equilibrada y realista desde el punto de vista fonético. Este aspecto es el que dio lugar a un mayor control sobre el diseño de la base de datos, sobre todo en el caso de los subcorpus de frases fonéticamente ricas y de números.

Las unidades de reconocimiento seleccionadas son alófonos y trifonemas. En la elección de los alófonos se seleccionaron aquellos que presentan una distribución complementaria en la lengua, y en la base de datos se aseguró un mínimo de 1000 repeticiones de cada uno de ellos. En total se consideraron necesarios para el reconocimiento 29 alófonos diferentes, entre los que se encuentran los 25 fonemas gallegos (7 vocálicos y 18 consonánticos), es decir, aquellos que tienen valor fonológico. En el caso de los fonemas /b/, /d/ y /g/ se incluyeron los alófonos complementarios oclusivos ([b], [d], [g]) y aproximantes ([B], [D], [G]) sonoros. También se incluyó el alófono fricativo velar sordo [x], con el fin de obtener un reconocimiento robusto en previsión de frecuentes nombres propios en castellano que pueden aparecer en los directorios de muchos telesistemas automáticos (lugares, apellidos...). Los alófonos considerados son los recogido en la Tabla 1.

La selección como unidad del trifonema tiene como objetivo representar a los alófonos de la lengua en todos sus contextos precedentes y siguientes posibles. Para seleccionar el conjunto de trifonemas relevantes, o sea, aquellos que se van a entrenar, en lugar de hacer un estudio teórico de las combinaciones fonemáticas posibles en gallego, se recurrió al análisis estadístico de las combinaciones que aparecen en un gran corpus textual. Este macrotexto recoge una gran cantidad de texto de carácter periodístico y su tamaño es lo suficientemente amplio como para presuponer que es representativo de la lengua. Los resultados nos sirvieron para establecer una comparación con el conjunto de elementos incluidos finalmente en la base de datos, con el fin de asegurar un corpus equilibrado en el que todos los alófonos y los trifonemas más significativos de la lengua tuviesen una representación mínima suficiente para el entrenamiento del reconocedor.

Los diez ítems reservados a las frases fonéticamente ricas constituyen el cuerpo fundamental de recogida de material fonético. Se estableció un umbral mínimo de 15 realizaciones para cada trifonema que se pretenda modelar. Así, teniendo en cuenta que la media de repeticiones de las frases fonéticamente ricas es de 3,75 veces, garantizamos un mínimo de 56 realizaciones de cada trifonema relevante, cantidad más que suficiente para el entrenamiento. El total de trifonemas se obtuvo con la inclusión de cuatro alocuciones con palabras fonéticamente ricas, con lo que se aumenta el número de contextos limitando con silencio.

3.2 El diseño de una base de datos de números naturales.

Los números naturales constituyen el otro elemento fundamental de la base de datos, ya que es uno de los aspectos del reconocimiento más difícil de tratar.

Una base de datos equilibrada de números no se consigue con el mismo número de realizaciones de cada número considerado, sino de cada modelo que se quiere entrenar. Así, en el diseño de la base de datos numérica se tuvieron en cuenta varios aspectos lingüísticos, como que muchos números se forman por composición o que algunos presentan alternancia de género, con el fin de elaborar un conjunto de datos compensado. Para ello se potenciaron especialmente los menos frecuentes, de la siguiente manera:

Clave	Alóf.	Caracterización	Ejemplo	Trans.
a	/a/	vocal central abierta	arpa	/arpa/
E	/E/	vocal anterior medio abierta	terra	/tERa/
e	/e/	vocal anterior medio cerrada	pena	/pena/
i	/i/	vocal anterior cerrada	mira	/mira/
O	/O/	vocal posterior medioabierta	cómpre	/kOmpRe/
o	/o/	vocal posterior mediocerrada	tolo	/tolo/
u	/u/	vocal posterior cerrada	furas	/furas/
p	/p/	oclusivo bilabial sordo	peito	/peito/
t	/t/	oclusivo dental sordo	teito	/teito/
k	/k/	oclusivo velar sordo	cana	/kana/
B	/B/	oclusivo bilabial sonoro aprox.	saba	/saBa/
b	/b/	oclusivo bilabial sonoro	cambio	/kambio/
D	/D/	oclusivo dental sonoro aprox.	medo	/meDo/
d	/d/	oclusivo dental sonoro	caldo	/kaldo/
G	/G/	oclusivo velar sonoro aprox.	pega	/peGa/
g	/g/	oclusivo velar sonoro	gota	/gota/
C	/tS/	africado palatal sordo	chover	/tSoBer/
m	/m/	nasal bilabial sonora	mesa	/mesa/
n	/n/	nasal alveolar sonora	noite	/noite/
J	/J/	nasal palatal sonora	poñer	/poJer/
N	/N/	nasal velar sonora	algunha	/alguNa/
f	/f/	fricativo labiodental sordo	ferreiro	/ferreiro/
T	/T/	fricativo interdental sordo	cear	/Tear/
s	/s/	fricativo alveolar sordo	sentar	/sentar/
S	/S/	fricativo prepalatal sordo	xunta	/Sunta/
Z	/Z/	lateral palatal sonora	coller	/koZer/
r	/r/	vibrante alveolar sonora	coroa	/koroa/
R	/rr/	vibrante múltiple alveolar sonora	carro	/kaRo/
x	/x/	fricativo velar sordo	Juan	/xuan/

Tabla 1: Lista de los alófonos del gallegos considerada en el diseño del material.

- Los números entre 10 y 19 frente a las unidades y, entre ellos, los contextos menos habituales: *dez, once, doce, trece, catorce, quince*.
- Los múltiplos exactos de diez: *vinte, trinta, corenta, cincuenta, sesenta, setenta, oitenta* y *noventa* (frente a *vinte, trinte-, corente-, cincuente-*..., que aparecen nueve veces más).
- Las unidades tipo *cen, cento, centos, mil, millón, millóns*, potenciando *cento* (frente a *centos*), *cen* (frente a *cento, centos*) y *millón* (frente a *millóns*), y teniendo en cuenta que las unidades más bajas son más recurrentes, pues tienen la posibilidad de aparición incluidas en los números más altos.
- Finalmente, se consideraron 41 modelos diferenciales: *cero, un, unha, dous, dúas, tres, catro, cinco, seis, sete, oito, nove, dez, once, doce, trece, catorce, quince, deza-, vinte, trinta, corenta, cincuenta, sesenta, setenta, oitenta, noventa, vinte-, trinte-, corente-, cincuente-, sesente-, oitente-, novente-, cen, cento, centos, centas, mil, millón, millóns.*

Otro aspecto lingüístico que se tuvo en cuenta fue el de la coarticulación. Se tuvo en cuenta que cada modelo apareciese en todas las posiciones numéricas posibles, por ejemplo, *centos* en la posición de centenas, de centenas de millar...; y en todos los contextos fonéticos anteriores y posteriores, por ejemplo, *centos* precedido en el margen anterior por -s (*dous, tres, seis*), -e (*sete, nove*) y por -o (*catro, cinco, oito*). Para posibilitar el reconocimiento de los comandos en un contexto de habla continua, uno de los ítems de la base de datos está reservado a frases con los comandos en 'word spotting'. El diseño de estas frases se hizo pensando en las posibles órdenes a un sistema automático, de forma que se entrenasen las palabras clave en los contextos más habituales. Por otro lado, el objetivo de obtener el máximo realismo posible implicó un especial control en el diseño de las cantidades de dinero, debido a la problemática que presenta el cambio inminente de moneda. Se recogieron tanto cantidades con pesetas como con euros, con un proporción algo mayor de las cantidades con euros. Se estudiaron detenidamente las posibles futuras maneras de expresar las cantidades con euros, dado que incluirán la novedad de los decimales. Las posibilidades previstas son muy variadas (*quince coma trinta euros, quince con trinta euros, quince e trinta euros, quince euros con trinta, quince euros e medio*, etc.) y en el diseño se tuvo especial cuidado en representar todos los contextos posibles en cada modalidad.

4 RECOGIDA DE LOS DATOS

La base de datos contempla una primera campaña de grabación de 1000 hablantes, a los que no se adiestra previamente con el fin de obtener un corpus oral que se acerque lo más posible a la realidad. Los informantes se seleccionan en función de los requisitos del SpeechDat y del objetivo de alcanzar una base de datos equilibrada también en la representación de la variedad de la población.

- Se procura una representación similar de los dos sexos.
- La edad de los informantes abarca desde los 16 a los 60 años. La distribución respeta los requisitos del SpeechDat, que obliga a incluir un 20% de hablantes con edades comprendidas entre los 16 y 30 años, otro 20% entre 31 y 45 años, y un 15% entre 46 y 60 años, pero se potencia de modo especial la inclusión del grupo comprendido entre los 20 y los 45 años.
- Otro aspecto que se tiene en cuenta es el nivel sociocultural de los informantes, potenciándose la grabación de hablantes del medio urbano y con una cierta formación académica, ya que se estima que es éste el grupo al que se ajusta en mayor medida el perfil de usuarios de teleservicios automáticos.
- Se atiende también a la representación de las grandes áreas dialectales del gallego, seleccionando informantes procedentes de diferentes zonas, emprendiendo la campaña en distintos puntos geográficos. Además, se controla la procedencia y el lugar de residencia de los informantes requiriéndoles esta información en una ficha de identificación.
- Un aspecto que se tiene también en cuenta es cuál es la lengua materna y cuál la lengua habitual de los informantes. Se controla así que haya una representación mayoritaria de 'buenos'hablantes, aunque también se contempla una cierta presencia de hablantes con un menor dominio del gallego, teniendo en cuenta que el gallego se encuentra en una situación de dos lenguas en contacto que provoca frecuentes interferencias.

La captación de locutores se hace por medio de una campaña en la que se distribuyen unas hojas informativas en las que se solicita la colaboración con el proyecto de investigación y se indican los objetivos de la recogida de llamadas. Como aliciente a la participación se anuncia el sorteo de dos premios de 25.000 pesetas entre los hablantes que completen las llamadas, y se adjuntan las bases legales del sorteo. Se especifican las condiciones de participación y se resalta la gratuidad de la chamada.

Para proceder a la grabación de los hablantes se diseñó el cuestionario que se distribuirá entre ellos. Este cuestionario contiene las indicaciones que orientan al usuario durante toda la llamada y el material concreto de grabación, con una serie de instrucciones prácticas relacionadas con el entorno de grabación, para eliminar en lo posible el ruido ambiental (apagando los electrodomésticos, evitando conversaciones simultáneas a la llamada...), y el material concreto de la base de datos. Por un lado se incluye una tabla con los elementos diseñados para la lectura, repartidos en una serie de 35 puntos. La tabla consta de dos columnas: la de la izquierda presenta el número de orden de los elementos y una pequeña indicación del contenido de cada ítem, y la de la derecha reproduce el material particular destinado a la lectura en cada sesión. En la parte superior de cada tabla se indica el código de la planilla, que forma parte de la base de datos numérica y sirve como número de identificación del usuario.

Por otra parte todos los cuestionarios presentan por escrito las preguntas diseñadas para obtener las realizaciones espontáneas previstas en la base de datos. Después se procede a la lectura del material de la tabla, y antes de finalizar la llamada, pero dándole a entender al hablante que las respuestas a las siguientes demandas sirven únicamente para recoger sus datos personales, se solicita la identificación del usuario con su nombre completo, DNI y número de teléfono.

5 ADQUISICIÓN Y ETIQUETADO

La grabación telefónica se realiza en un entorno no controlado, por lo general la casa del informante, aunque procurando evitar en lo posible en ruido ambiental. El informante mantiene un diálogo con un sistema automático que le orienta durante toda la grabación y adquiere el material después de un tono que sigue a la correspondiente indicación del sistema. La señal sonora se adquiere directamente en formato digital por medio de un programa de adquisición automática de bases de datos, y se presenta en un directorio que reproduce el esquema organizativo y la denominación estándar de los ficheros SpeechDat. El nombre de los archivos proporciona información sobre el tipo de base de datos (recogida desde la red fija, en idioma gallego, en ficheros de sonido á-law'), y sobre el material recogido (número de la llamada y código del ítem con el que se corresponde la realización).

Finalmente, una vez recogidos los datos orales, se procede a su etiquetado. La representación escogida es la ortográfica, ya que es la convención adoptada por el proyecto SpeechDat. El inventario de etiquetas incluye unidades de anotación lingüística y paralingüística, es decir, se procede a la transliteración de las realizaciones orales, pero también se marcan los ruidos, tanto ambientales (portazos, timbre, música de fondo...) como propios del hablante (pronunciaciones ininteligibles o parciales, y también hesitaciones, pausas, clicks, risas...).

6 CONCLUSIONES

El diseño de la base de datos SpeechDat recogida para el gallego se realizó con un control exhaustivo de todo el material, tanto desde el punto de vista del respeto riguroso de las especificaciones del proyecto SpeechDat, como desde la perspectiva de obtener una base de datos compensada fonéticamente, en la que las unidades de entrenamiento del reconocedor tienen una presencia suficiente y equilibrada. Además, se aumentaron o adaptaron los elementos de la base de datos en función de las características propias del gallego no contempladas en el proyecto general, y de su adaptabilidad a todo tipo de teleservicios automáticos. Finalmente, se procuró la mayor naturalidad posible en las realizaciones orales. Así, en la selección del material se buscó el máximo realismo, con la inclusión de contenidos reales y frecuentes en la lengua, y se evitó en lo posible el condicionamiento de los hablantes derivados de la presentación escrita de los datos y del propio contexto de grabación.

Referencias

[1] C. Draxler G. Richard K. Choukri S. Euler J. Allen A. Moreno, B. Lindberg. SPEECH DAT CAR. A Large Speech Database For Automotive Environments. *Proceedings of the Second International Conference on Language Resources and Evaluation. Athens (Greece)*, 2000.

[2] J. van Velden F. Senia. Specification of orthographic transcription and lexicon conventions . *Technical Report. Project SpeechDat LE2-4001*, 1997.

[3] TELEMATICS. proyecto SPEECH DAT CAR. http:www.speechdat.org.

[4] TELEMATICS. proyecto SPEECH DAT M num. 63314. http:www.speechdat.org .

[5] TELEMATICS. Proyecto SPEECH DAT II num. LE2-4001. http:www.speechdat.org. 2000.

[6] R. Winski. Definition of corpus, scripts and standars for Fixed Networks . *Technical Report. Project SpeechDat LE2-4001*, 1997.