Llisterri, J., Machuca, M. J., de la Mota, C., Riera, M., & Ríos, A. (2005). Corpus orales para el desarrollo de las tecnologías del habla en español. *Oralia. Análisis del Discurso Oral, 8*, 289-325.

http://liceu.uab.cat/~joaquim/publicacions/Llisterri_Machuca_Mota_Riera_Rios_05_Corpus_Orales_Tecnologias_Habla_Espanol.pdf

CORPUS ORALES PARA EL DESARROLLO DE LAS TECNOLOGÍAS DEL HABLA EN ESPAÑOL

JOAQUIM LLISTERRI, MARÍA J. MACHUCA, CARME DE LA MOTA, MONTSERRAT RIERA Y ANTONIO RÍOS
Universitat Autònoma de Barcelona

1. INTRODUCCIÓN

La noción de corpus oral recubre muy diversos tipos de recursos lingüísticos, cuyo denominador común es que todos ellos reflejan, de un modo más o menos directo, la lengua hablada. Sin embargo, los objetivos de la investigación o del desarrollo que se pretende abordar una vez constituido el corpus, por una parte, y las diferencias de tradición académica entre las humanidades y las disciplinas tecnológicas, por otra, tienen como consecuencia que tanto se considere un corpus oral a un conjunto de transcripciones ortográficas –más o menos enriquecidas con anotaciones– de entrevistas entre un investigador y su informante o de programas de radio y televisión, como a una grabación de series de palabras aisladas, de números o de frases fonéticamente controladas realizada en el interior de un vehículo en marcha y a través de un teléfono móvil. Por este motivo se ha recurrido, en ocasiones, a la distinción entre corpus de lengua oral (*spoken language corpora*) y corpus orales (*speech corpora*): mientras que los primeros son, esencialmente, transcripciones de producciones lingüísticas más o menos espontáneas, los segundos tienen como núcleo la señal sonora y suelen ofrecer materiales mucho más controlados (Llisterri, 1996).

En este trabajo nos centraremos en los corpus orales en español diseñados y recogidos para desarrollar aplicaciones en el ámbito de las tecnologías del habla. Se trata, por ello, de recursos cuyo objetivo no es tanto la investigación lingüística –aunque no cabe duda de que algunos de ellos pueden llegar a ser sumamente útiles, especialmente en los estudios fonéticos– sino la creación de sistemas de conversión de texto en habla, de reconocimiento automático del habla o de diálogo entre personas y ordenadores.

En primer lugar, presentaremos brevemente algunos de los recursos disponibles en español, prestando especial atención a aquellos que pueden obtenerse a través de las agencias de distribución que se mencionan más adelante. La segunda parte de nuestra contribución se centrará en la transcrip-

ORALIA, Vol. 8, 2005, págs. 289-325.

ción y anotación fonética de corpus orales, y en ella expondremos, de un modo necesariamente sucinto, los principales sistemas desarrollados, así como algunas de las herramientas de dominio público que pueden emplearse para el análisis.

2. Los corpus orales para el desarrollo de las tecnologías del habla

Una primera característica que debe señalarse en lo que se refiere a los corpus orales en el ámbito de las tecnologías del habla es que, dada su aplicación, algunos de estos recursos han sido desarrollados por empresas o para empresas, por lo que no se trata, en general, de materiales fácilmente accesibles. En otros casos, los corpus se han creado en el marco de proyectos de I+D que han contado con fuentes de financiación pública españolas o europeas y debería ser posible, en estas situaciones, alcanzar acuerdos para su uso sin fines comerciales en otros proyectos de investigación. Una tercera categoría la constituyen los corpus que, con independencia de su origen, se distribuyen –en ocasiones con precios diferentes para empresas y para centros de investigación- a través de dos organismos: ELDA (*Evaluations and Language resources Distribution Agency*) en Europa y LDC (*Linguistic Data Consortium*) en Estados Unidos. En el presente trabajo nos referiremos principalmente a los recursos que se originaron a partir de proyectos de investigación con financiación pública y a los que se encuentran en los catálogos de las dos agencias mencionadas.

En segundo lugar, cabe tener en cuenta que para el desarrollo de las tecnologías del habla es casi imprescindible, salvo en algunos casos muy concretos, disponer de la señal sonora, a diferencia de lo que sucede en otro tipo de estudios lingüísticos que pueden llevarse a cabo a partir de transcripciones ortográficas de la lengua oral. En este sentido, los corpus que tratamos aquí tienen una característica común con el tipo de recursos que se necesitan para los estudios fonéticos: la señal sonora, adecuadamente segmentada y etiquetada, constituye una parte esencial del corpus.

En los apartados que siguen presentamos, clasificados según su principal aplicación, algunos de los corpus orales con los que contamos en español para abordar proyectos relacionados con las tecnologías del habla. La información se resume en forma de tablas, en las que se indica sucintamente el contenido lingüístico del corpus, el número de locutores grabados, el canal a través del cual se han recogido los datos y, finalmente, la disponibilidad; en este caso, se señala, si procede, el número de catálogo de ELDA o del LDC, o la institución responsable del desarrollo del corpus; debe tenerse en cuenta que se incluyen también corpus multilingües en los que está presente el español en cualquiera de sus variantes geográficas. Otros recursos desarrollados en el marco de diversos proyectos, pero de los que no tenemos

ninguna indicación sobre su accesibilidad se mencionan en el texto del trabajo, sin incluirlos en las tablas. Debemos señalar, finalmente que, por razones de espacio, para cada uno de los proyectos y recursos hemos intentado seleccionar únicamente una referencia bibliográfica, pese a que algunos de ellos están abundamente documentados; en general, hemos procurado que se tratara de la descripción más reciente o de la más accesible.

2.1. *Corpus generales para el desarrollo de aplicaciones en tecnologías del habla*

No suele ser habitual, en la actualidad, encontrar recursos de tipo general para lenguas como el español que faciliten el desarrollo de aplicaciones en el campo de las tecnologías del habla. Sin embargo, en determinadas circunstancias, puede ser útil disponer de un corpus con información básica y, por ello, en el marco de los proyectos SAM (*Multilingual Speech Input/Output Assessment, Methodology and Standardisation*) y SAM-A (*Speech Technology Assessment in Multilingual Applications*), llevados a cabo entre 1986 y 1995, se creó el corpus multilingüe EUROM (Chan *et alii,* 1995). La importancia de SAM y de EUROM radica también en que contribuyeron a definir un conjunto de estándares como parte del proyecto EAGLES (*Expert Advisory Group on Language Engineering Standards*) (Gibbon *et alii,* 1997) que, en el contexto europeo, siguen empleándose en la actualidad.

	Contenido	**Locutores**	**Canal**	**Disponibilidad**
EUROM EUROM1 The multilingual European speech database (Chan *et alii,* 1995)	Combinaciones C(C)VC(V) aisladas y en contexto (5 pares de palabras en el contexto) 100 números que recogen las posibilidades fonotácticas de la lengua 40 párrafos de 5 frases cada uno 50 frases para complementar el equilibro fonético de los párrafos	30 locutores masculinos y 30 locutores femeninos por lengua	Grabación en estudio	ELDA-S0014

TABLA 1: Corpus generales para el desarrollo de tecnologías del habla.

2.2. *Corpus para el desarrollo de sistemas de conversión de texto en habla*

El desarrollo de un sistema de conversión de texto en habla requiere disponer de corpus que contengan datos para extraer las unidades que se em-

plearán en la síntesis y para definir los modelos prosódicos que utilizará el conversor (Llisterri *et alii,* 2004a). Los corpus creados para tal fin suelen recoger los enunciados de un único locutor o de un número reducido de hablantes, puesto que el objetivo que se persigue es reproducir, de la manera más natural posible, el modo en que una persona leería un texto en voz alta. Buena parte de los corpus diseñados para esta aplicación se han creado teniendo en cuenta un sistema específico ligado a las necesidades de una empresa (véase, por ejemplo, Llisterri *et alii,* 2004b) o de un grupo de investigación, por lo que, como se observa en la tabla 2, no se dispone de un gran número de recursos fácilmente accesibles.

	Contenido	**Locutores**	**Canal**	**Disponibilidad**
Multext Prosodic Database (Campione y Véronis 1998)	Párrafos leídos del corpus EUROM Estilización de la curva de F_0 y anotación mediante INTSINT (véase el apartado 3.2.1.) Transcripción ortográfica alineada con la señal	5 locutores masculinos y 5 locutores femeninos por lengua	Grabación en estudio	ELDA-S0060
Spanish TTS Speech Corpus (Appen)	1787 frases que recogen los difonos y trifonos de la lengua Léxico de 3748 palabras transcritas	1 locutor masculino	Grabación en estudio	ELDA-S0150

TABLA 2: Corpus para el desarrollo de sistemas de conversión de texto en habla.

Dado que un conversor de texto en habla debe estar adaptado a la variante geográfica de sus potenciales usuarios, se han diseñado corpus para las distintas modalidades del español de América (Renato y Álvarez, 2004) o para variedades específicas como las de Argentina (Gurlekian *et alii,* 2001b) o de México (*Tlatoa/OGI Spanish TTS Corpus*). Existen, además, bases de datos bilingües como la que se presenta para el castellano y el catalán en Esquerra *et alii* (1998).

La posibilidad de dotar de mayor flexibilidad a la síntesis ha hecho que algunas investigaciones en este campo se hayan orientado hacia el estudio de las emociones en el habla, con objeto de modelarlas en los sistemas automáticos. Algunos proyectos llevados a cabo para el español han recurrido a bases de datos ya existentes (Iriondo *et alii,* 2000; Rodríguez *et alii,* 1999), mientras que en otros casos se han constituido recursos específicos; así, en el marco del proyecto VAESS *(Voices, Attitudes and Emotions in Speech Synthesis)*

se desarrolló la base de datos conocida como SES (*Spanish Emotional Speech*), grabada por un actor profesional, en la que se recogen emociones como la tristeza, la alegría, el enfado o la sorpresa mediante la lectura de textos breves (Montero *et alii,* 1998). Más recientemente, en el proyecto INTERFACE (*Multimodal Analysis/Synthesis System for Human Interaction to Virtual and Augmented Environments*) se ha recogido un corpus multimodal que incluye el español, grabado por un actor masculino y por otro femenino, reproduciendo seis emociones –enfado, tristeza, alegría, miedo, disgusto y sorpresa– en palabras aisladas, en frases de diferente duración y en textos (Hozjan *et alii,* 2002).

2.3. *Corpus para el desarrollo y la evaluación de sistemas de reconocimiento del habla*

Contrariamente a lo que sucede en el caso de la síntesis, los corpus orientados hacia el reconocimiento del habla recogen producciones lingüísticas de un elevado número de locutores. La razón es que el desarrollo de un sistema de reconocimiento requiere una etapa de entrenamiento, en la que se emplea un corpus oral para crear los modelos acústicos de las unidades que se utilizarán en el reconocimiento (Llisterri *et alii,* 2003; Mariño y Nadeu, 2004). Un mayor número de locutores, de variantes geográficas y de estilos de habla permiten elaborar modelos acústicos que reflejen de un modo más adecuado la realidad fonética de una lengua. En una segunda fase del desarrollo, el sistema se evalúa mediante una parte del corpus que no se haya empleado en el entrenamiento, lo que hace necesario disponer de una cantidad adicional de datos.

De un modo análogo, se han empleado corpus, como los denominados HUB-4 y HUB-5 (véase la tabla 3), para comparar los resultados de distintos sistemas de reconocimiento de habla con un mismo conjunto de datos; estas evaluaciones han sido llevadas a cabo principalmente por el NIST (*National Institute of Standards and Technology*) en Estados Unidos y se encuentran documentadas en las páginas en Internet de la institución.

El contenido lingüístico de los corpus para el reconocimiento del habla presenta también sus especificidades. Como se aprecia en la tabla 3, se incorporan palabras directamente relacionadas con el dominio para el que se desarrollará una aplicación; así, podemos encontrar números de tarjetas de crédito para la banca telefónica, o topónimos y antropónimos para servicios de información. Es interesante observar que algunas bases de datos incluyen conjuntos de frases fonéticamente equilibradas –es decir, con una frecuencia de aparición de alófonos o de fonemas que refleja la que es general en la lengua– o fonéticamente ricas –preparadas para que, en el momento

de entrenar el reconocedor, se encuentre un número suficiente de muestras de cada alófono o fonema de la lengua–, que pueden ser útiles en estudios fonéticos, junto con los fragmentos de habla semiespontánea que se recogen en algunos corpus.

Por otra parte, un sistema de reconocimiento incluye entre sus componentes lo que se conoce como modelo de lenguaje, consistente, en esencia, en probabilidades de aparición de combinaciones de palabras en función del contexto; para recopilar esta información se recurre a transcripciones ortográficas de la lengua oral que, en algunas ocasiones, pueden corresponder a los enunciados que forman parte del corpus empleado para modelar las unidades acústicas.

Este tipo de corpus suele distribuirse también acompañado de un diccionario de pronunciación, en el que se asocia la representación ortográfica con la transcripción fonética del léxico que aparece en el corpus; ésta se realiza mediante un alfabeto fonético computacional (véase el apartado 3.1) y recoge, habitualmente, todas las variantes documentadas en el corpus.

Una última característica de los recursos que describimos en este apartado es que, en muchos casos, se adquieren en el entorno en que tendrá que utilizarse el sistema de reconocimiento; esto se debe a la necesidad de realizar un entrenamiento y una evaluación en condiciones reales y, por tal motivo, se han recogido bases de datos a través de la red de telefonía fija y móvil y en diversas situaciones –en una oficina, en la calle, en el interior de un vehículo en marcha– en las que un potencial usuario podría acceder a un servicio basado en el reconocimiento del habla.

TABLA 3: Corpus para el desarrollo de sistemas de reconocimiento del habla.

	Contenido	**Locutores**	**Canal**	**Disponibilidad**
Albayzín (Casacuberta *et alii*, 1992)	Corpus fonético: 6.800 frases fonéticamente equilibradas Corpus geográfico: 6800 frases de consulta a una base de datos geográfica Corpus Lombard: 200 frases grabadas con efecto Lombard[1]	152 locutores masculinos 152 locutores femeninos	Grabación en estudio	ELDA-S0089

[1] En estos casos, el locutor escucha, a través de unos auriculares, un ruido de fondo, de modo que su producción resulta similar a la que se daría en un ambiente ruidoso.

TABLA 3: Corpus para el desarrollo de sistemas de reconocimiento del habla *(continuación)*.

	Contenido	**Locutores**	**Canal**	**Disponibilidad**
ANITA Audio eNhancement In secured Telecom Applications	60 frases fonéticamente ricas Letras y números En situación normal y en situación de estrés y de pánico 10 minutos de texto literario leído en situación normal	17 locutores masculinos 24 locutores femeninos	Micrófono En varios entornos con diferente ruido ambiental	ELDA-S0156
AURORA Subset of SpeechDat-Car Spanish	Dígitos aislados Dígitos conectados		Micrófono En el interior de un coche con diferentes condiciones de ruido	ELDA AURORA CD0003-02
CALLHOME Spanish Speech CALLHOME Spanish Transcripts CALLHOME Spanish Lexicon CALLHOME Spanish Dialogue Act Annotation	120 conversaciones telefónicas con una duración máxima de 30 minutos cada una Transcripción de segmentos de entre 5 y 10 minutos de duración Léxico de 45.582 palabras Anotación de la estructura del diálogo		Teléfono fijo	LDC96S35 LDC96T17 LDC2001T61 LDC96L16
HUB-4 Broadcast News Evaluation 1997 Non English Test Material	1 hora de grabaciones de noticias procedentes de Televisa y Univisión		Medios de comunicación	LDC2001S91
HUB-4 NE Spanish Broadcast News Speech 1997 HUB-4 NE Spanish Broadcast News Transcripts 1997	30 horas de noticias procedentes de Televisa, Univisión y VOA Transcripción ortográfica de las grabaciones		Medios de comunicación	LDC98S74 LDC98T29

TABLA 3: Corpus para el desarrollo de sistemas de reconocimiento del habla *(continuación)*.

	Contenido	**Locutores**	**Canal**	**Disponibilidad**
HUB-5 Spanish Evaluation HUB-5 Spanish Transcripts	Conversaciones telefónicas Transcripción ortográfica de 20 conversaciones		Teléfono fijo	LDC2002S25 LDC2003T04
HUB-5 Spanish Telephone Speech Corpus HUB-5 Spanish Transcripts	106 llamadas de 10 a 30 minutos de duración extraídas del corpus CallFriend Transcripción ortográfica de las grabaciones		Teléfono fijo	LDC98S70 LDC98T27
LATINO-40 Spanish Read News	5000 frases leídas, seleccionadas de textos extraídos de periódicos	20 locutores masculinos 20 locutores femeninos	Grabación en estudio	LDC95S28
MICROADES, ATLAS Spanish Microphone Database	450 párrafos Transcripción ortográfica Léxico de 7400 palabras con la transcripción en SAMPA (véase el apartado 3.1.1.)	300 locutores	Grabación en estudio 4 micrófonos diferentes	ELDA-S0165
SALA II Spanish from Mexico database	Estándares definidos en el proyecto SALA	539 locutores masculinos 536 locutores femeninos	Teléfono móvil En varios entornos con diferente ruido ambiental	ELDA- S0171
SALA II Spanish Mobile Network Database collected in Venezuela	Estándares definidos en el proyecto SALA	576 hablantes masculinos 603 hablantes femeninos	Teléfono móvil En varios entornos con diferente ruido ambiental	ELDA-S0167
SALA Spanish Colombian Database	Estándares definidos en el proyecto SALA	475 hablantes masculinos 525 hablantes femeninos	Teléfono fijo	ELDA-S0084
SALA Spanish Venezuelan Database	Estándares definidos en el proyecto SALA	504 hablantes masculinos 496 hablantes femeninos	Teléfono fijo	ELDA-S0141

TABLA 3: Corpus para el desarrollo de sistemas de reconocimiento del habla *(continuación)*.

	Contenido	Locutores	Canal	Disponibilidad
Siemens Chilean Spanish FDB-500	12168 enunciados Dígitos y palabras relacionadas con la aplicación Léxico con la transcripción en SAMPA	235 locutores femeninos 272 locutores masculinos	Teléfono fijo	ELDA-S0054
Siemens Colombian Spanish Speech Database	Dígitos y palabras relacionadas con la aplicación	563 locutores masculinos 502 locutores femeninos	Teléfono fijo	ELDA-S0064
Spanish Speech Corpus 1 (Appen)	100 palabras para el control de aplicaciones 100 frases fonéticamente ricas Léxico de 3748 palabras transcritas mediante SAMPA	100 locutores masculinos 100 locutores femeninos	Micrófono En distintas condiciones de ruido ambiental	ELDA-S0149
Spanish Speecon database (Iskra *et alii* 2002)	5 minutos de habla espontánea (adultos) Lectura de 30 (adultos) o 60 (niños) frases y de 5 palabras fonéticamente ricas Palabras y números relacionados con aplicaciones (adultos) o con el control de juegos (niños) Léxico transcrito mediante SAMPA	279 locutores masculinos adultos 282 locutores femeninos adultos 27 locutores masculinos entre 8 y 14 años 28 locutores femeninos entre 8 y 14 años	Micrófono En entornos con diferente ruido ambiental	ELDA-S0160
SpeechDat Spanish (II) FDB-1000 SpeechDat Spanish (II) FDB-4000	Estándares definidos en el proyecto SpeechDat	481 locutores masculinos 519 locutores femeninos 2061 locutores masculinos 1939 locutores femeninos	Teléfono fijo	ELDA-S0101 ELDA-S0102
SpeechDat Spanish (M) DB1 SpeechDat Spanish (M) DB2	Estándares definidos en el proyecto SpeechDat	508 locutores masculinos 494 locutores femeninos	Teléfono fijo	ELDA-S0065 ELDA-S0066

TABLA 3: Corpus para el desarrollo de sistemas de reconocimiento del habla *(continuación)*.

	Contenido	**Locutores**	**Canal**	**Disponibilidad**
SpeechDat Spanish Database for the Mobile Telephone Network	Estándares definidos en el proyecto SpeechDat	526 locutores masculinos 540 locutores femeninos	Teléfono móvil	ELDA-S0119
SpeechDat-Car Spanish Database	Estándares definidos en el proyecto SpeechDat	156 locutores masculinos 150 locutores femeninos	Micrófono y teléfono móvil En el interior de un coche	ELDA-S0140
VAHA, Voice Across Hispanic America (Polyphone II)	39000 enunciados	915 locutores	Teléfono fijo	LDC96S41

Es interesante señalar que, entre los recursos disponibles mencionados en la tabla 3, Albayzín (Casacuberta *et alii,* 1992) constituyó un proyecto en el que participaron seis grupos de investigación españoles con el objetivo de crear un recurso común útil para el entrenamiento y la evaluación de sistemas de reconocimiento de habla; en él se prestó especial atención, entre otros aspectos, al diseño del contenido fonético del corpus (Moreno *et alii,* 1993).

Entre las iniciativas llevadas a cabo en el contexto europeo destacaríamos los sucesivos proyectos SpeechDat, iniciados en 1995: SpeechDat (*Infrastructure for Spoken Language Resources*), SpeechDat *(Databases for the Creation of Voice Driven Teleservices)* (Draxler *et alii,* 1998) y SpeechDat-Car (*Speech Databases recorded in Vehicles*) (Moreno *et alii,* 2000), a los que han seguido otros centrados ya en lenguas eslavas o semíticas. En los tres citados se han creado recursos para el español, y los estándares definidos para diseñar y recoger corpus dedicados al reconocimiento del habla se han aplicado en muchas ocasiones a otros proyectos. En lo que se refiere al contenido lingüístico, las bases de datos definidas en el proyecto SpeechDat comprenden: un dígito aislado, 4 dígitos conectados (por ejemplo, números de teléfono o de tarjetas de crédito), 2 números naturales, un número con decimales, 2 cantidades de dinero, 3 palabras deletreadas, 2 expresiones de tiempo, 3 fechas, 3 respuestas a preguntas absolutas, un topónimo, 6 palabras relacionadas con aplicaciones del reconocimiento, 3 palabras relacionadas con aplicaciones incluidas en frases y 9 frases leídas para asegurar la cobertura fonética de la base de datos. En ocasiones se añaden otros elementos, como palabras rela-

cionadas con el entorno en el que se recogió el corpus o con usos de la telefonía móvil. Las bases de datos incorporan, además, un léxico transcrito mediante SAMPA (véase el apartado 3.1) en el que se recoge el vocabulario que aparece en el corpus (Winski, 1997; Velden *et alii,* 1996).

Cabe mencionar también SALA (*SpeechDat acrosss Latin America*) que, a través de SALA I (*Fixed telephone nework - Latin America*) (Moreno *et alii,* 2000) y SALA II (*Mobile/Cellular telephone network - Latin America, US & Canada*) (Heuvel *et alii,* 2004), ha contribuido a que, en la actualidad, se disponga de corpus para el desarrollo de sistemas de reconocimiento específicamente adaptados a las distintas variantes del español de América, tal como puede observarse en la tabla 3[2]. El contenido lingüístico de las bases de datos del proyecto SALA responde, por lo general, a los estándares de SpeechDat, del que se considera una extensión, y comprende los siguientes elementos: 6 palabras relacionadas con aplicaciones del reconocimiento, una secuencia de 10 dígitos aislados, 4 dígitos conectados (por ejemplo, números de teléfono, de tarjetas de crédito o de códigos PIN), 3 fechas, 1 frase que contiene alguna de las palabras relacionadas con la aplicación, un dígito aislado, 3 palabras deletreadas, una cantidad expresada con el nombre de la moneda, un número natural, cinco nombres propios (antropónimos, topónimos y nombres de empresas), 2 respuestas a preguntas absolutas, 2 expresiones temporales, 9 frases fonéticamente ricas y 4 palabras fonéticamente ricas (Moreno *et alii,* 2002).

En el contexto de los trabajos de grupos de investigación dedicados al reconocimiento del habla se han desarrollado otros corpus para el español de Argentina (Gurlekian *et alii,* 2001) siguiendo, en este caso, las especificaciones de SALA, y para el de México (*Tlatoa Common Questions Corpus*; Uraga y Gamboa, 2004; Villaseñor *et alii,* 2004); en algunos recursos se ha cuidado especialmente el contenido lingüístico, como sucede en DIMEx100 (Pineda *et alii,* 2004), realizado en colaboración con expertos en fonética del español de México. Por su parte, el interés comercial del reconocimiento ha llevado a empresas como Siemens a crear bases de datos que también se distribuyen a través de ELDA (véase la tabla 3) o a otras, como Telefónica I+D, a constituir sus propios recursos (Esteve *et alii,* 1994).

---

[2] A modo de ejemplo, la última base de datos incorporada al catálogo de ELDA (la del español de México elaborada en el marco del proyecto SALA II) se distribuye, si el comprador es socio de ELDA, a un precio de 34.000 euros si se emplea para la investigación y de 45.000 euros para usos comerciales; para los compradores que no son socios de ELDA, el precio es de 40.000 o de 51.000 euros respectivamente. El corpus del español de Venezuela puede adquirirse, en cambio, por un precio que oscila entre los 20.000 y los 30.00 euros en el caso de los datos recogidos por teléfono móvil (SALA II) y entre 13.000 y 20.000 euros los adquiridos a través de la red de telefonía fija (SALA I).

Una de las tendencias más recientes en el campo de los corpus orales es la creación de recursos multimodales que incorporan simultáneamente voz e imagen; aunque este tipo de corpus encuentra su principal aplicación en el diseño de sistemas de diálogo (véase el apartado 2.6.), pueden emplearse para el reconocimiento del habla; éste es el caso de AV@CAR, un corpus multimodal para desarrollar un sistema de reconocimiento de habla en español que, complementado con información visual, pueda utilizarse en vehículos (Ortega *et alii,* 2004).

2.4. *Corpus para la identificación automática del locutor*

Existen, hoy en día, servicios, como la banca telefónica, que podrían ser más accesibles si la verificación de la identidad del usuario se realizara a través de su propia voz; por otra parte, en el contexto judicial y en el de la seguridad se plantea también el uso de la voz de una persona como procedimiento para su identificación. Por este motivo, se han recopilado corpus orales que tienen como objetivo permitir el entrenamiento de sistemas automáticos de identificación y verificación del locutor.

En español contamos principalmente con el corpus Ahumada (Ortega *et alii,* 2000), que contiene dígitos, frases fonéticamente equilibradas, un texto fonéticamente equilibrado leído a tres velocidades diferentes y habla semiespontánea en forma de narración o de descripción de una imagen. Las grabaciones de 224 locutores masculinos y 231 locutores femeninos se realizaron en estudio y a través del teléfono, en tres sesiones espaciadas en el tiempo, un rasgo esencial en este tipo de corpus. Ahumada se está utilizando también en los análisis fonéticos del proyecto VILE, dedicado al análisis acústico de la variación inter e intra locutor en español (Battaner *et alii,* 2003).

En el ámbito que nos ocupa, cabe mencionar también TelVoice (Rodríguez *et alii,* 2003), con 59 locutores grabados a través del teléfono; Polycost (Hennebert *et alii,* 2000), un corpus para el reconocimiento del hablante recogido por el mismo medio, que incluye el español, y el reciente corpus telefónico multilingüe Mixer (Cieri *et alii,* 2004), que presenta la particularidad de incorporar hablantes bilingües inglés-español.

2.5. *Corpus para la identificación automática de la lengua*

Otra de las necesidades surgidas de la implantación de nuevos servicios telefónicos es la identificación automática de la lengua en la que se expresa un usuario. Desarrollar este tipo de tecnologías requiere también disponer de corpus que contengan, en este caso, una amplia variedad de lenguas. En

la tabla 4 se muestran algunos de los recursos disponibles que incluyen el español, a los que podría añadirse el corpus descrito por Lamel *et alii* (1998).

TABLA 4: Corpus para el desarrollo de sistemas de identificación automática de la lengua.

	Contenido	**Canal**	**Disponibilidad**
22 Language Corpus v1.2. (Lander *et alii* 1995)	Respuestas a preguntas concretas sobre datos personales de los locutores Habla espontánea	Teléfono	Center for Spoken Language Understanding, Oregon Graduate Institute
CALLFRIEND Spanish-Caribbean Dialect CALLFRIEND Spanish-Non-Caribbean Dialect	60 conversaciones telefónicas de 5 a 30 minutos de duración	Teléfono	LDC96S57 LDC96S58
Multilanguage Telephone Speech Corpus v1.2. OGI Multilanguage Corpus (Muthusamy *et alii*, 1992)	Palabras aisladas 108 muestras de habla espontánea de 1 minuto cada una transcritas fonéticamente	Teléfono	Center for Spoken Language Understanding, Oregon Graduate Institute LDC94S17

2.6. *Corpus para el desarrollo de sistemas de diálogo*

El uso cada vez más creciente de sistemas de diálogo mediante los que una persona puede acceder a un servicio telefónico automático para obtener un determinado tipo de información o realizar una transacción hace que cuestiones tales como la metodología del diseño de sistemas conversacionales, la constitución de recursos para desarrollarlos o la estandarización de los sistemas de anotación de corpus de diálogo sean unas de las más abordadas en la actualidad en el campo de las tecnologías del habla.

Uno de los principales problemas al que se enfrenta el investigador es que si bien, como acabamos de constatar, existen corpus útiles para desarrollar sistemas de síntesis y de reconocimiento, la creación de un sistema de diálogo suele responder a necesidades tan específicamente relacionadas con su aplicación futura que es difícil plantearse la reutilización de recursos ya constituidos a la hora de abordar un nuevo proyecto.

En el desarrollo de un sistema de diálogo pueden distinguirse dos tipos de corpus: el denominado «persona-persona», que intenta recoger una muestra amplia y realista de la situación comunicativa propia de la aplicación, y el corpus «persona-máquina», que contiene las intervenciones entre potenciales usuarios y un prototipo o una simulación del sistema (Llisterri *et alii*, 2003).

El corpus «persona-persona» es necesario para determinar los conocimientos que deben incorporarse al sistema y para establecer las futuras estrategias de gestión del diálogo en función del comportamiento natural observado en los operadores y en los usuarios de un determinado servicio. En cambio, mediante un corpus «persona-máquina» se pretende conocer la reacción de los futuros clientes de un servicio, así como detectar los errores del prototipo del sistema de diálogo que se está desarrollando.

Por esta razón, los corpus «persona-máquina» se recogen mediante un protocolo conocido como «el Mago de Oz», en el cual el denominado «mago» es una persona especialmente formada que realiza las funciones propias del sistema automático sin que el hablante sepa que, en realidad, no se está comunicando con un ordenador. Se trata, pues, de interacciones ficticias en las que los participantes en la recogida del corpus deben seguir las pautas marcadas en un conjunto de escenarios correspondientes a las futuras funcionalidades del sistema.

En algunas ocasiones se recogen también corpus con diálogos entre personas y una versión del sistema ya en funcionamiento; con ello se obtienen, por ejemplo, datos que permiten mejorar el módulo de reconocimiento incorporando los fenómenos propios del habla espontánea o, incluso, diseñando estrategias para corregir, como un paso previo al reconocimiento, las llamadas «disfluencias» de la lengua oral.

Los corpus de diálogo suelen anotarse teniendo en cuenta otros niveles –especialmente los relacionados con la pragmática– además del fonético segmental y suprasegmental, habitual en los recursos que hemos presentado hasta ahora. Uno de los problemas que se plantean en estos momentos es la multiplicidad de esquemas de anotación, tal como puede verse en la revisión de Klein *et alii* (1998), llevada a cabo en el marco del proyecto MATE (*Multilevel Annotation, Tools Engineering*). Para la anotación de corpus multimodales se cuenta, igualmente, con una importante diversidad de esquemas, que se presentan en Wegener *et alii* (2002), resultado del trabajo del grupo sobre interacción natural y multimodalidad de la iniciativa ISLE (*Internacional Standards for Language Engineering*), y en Serenari *et alii* (2002), específicamente centrado en la codificación de gestos y expresiones faciales, una de las cuestiones abordadas en el proyecto NITE (*Natural Interactivity Tools Engineering*). Cabe mencionar que se han diseñado algunas propuestas de anotación basadas en corpus en español, como las de Martínez *et alii* (2002) para BASURDE o de Villaseñor *et alii* (2000) para DIME, recursos que se describen a continuación.

Existen, indudablemente, en España grupos de investigación y empresas que han desarrollado sistemas de diálogo (Llisterri, 2003), pero la documentación disponible se centra más en la descripción de las aplicaciones o en las tareas realizadas por los distintos módulos del sistema que en los re-

cursos creados para llegar al resultado final. Podemos, sin embargo, mencionar el corpus «persona-persona» recogido en el proyecto BASURDE (Desarrollo de un sistema de diálogo oral el dominios restringidos) (Bonafonte *et alii,* 2000), consistente en 204 diálogos en el dominio de la información sobre viajes en ferrocarril; la transcripción ortográfica con algunas marcas de codificación está disponible en las páginas web del proyecto. En el contexto del proyecto DIHANA (Sistema de diálogo para el acceso a la información mediante habla espontánea en diferentes entornos) que, en cierto modo, constituye la continuación de BASURDE, se ha descrito el procedimiento de adquisición del corpus mediante la técnica del Mago de Oz (Alcácer *et alii,* 2004).

Para el español de México se cuenta con el corpus DIME (Diálogos Inteligentes Multimodales en Español), un recurso multimodal recogido mediante el protocolo del Mago de Oz, que incluye tanto la señal sonora como la filmación de las acciones que tienen lugar en la pantalla de un ordenador en un entorno para el diseño de cocinas (Pineda *et alii,* 2002).

2.7. *Corpus para la traducción automática del habla*

Uno de los retos actuales en el campo de las tecnologías del habla es lograr la traducción automática y en tiempo real de conversaciones entre personas que emplean lenguas diferentes. Para desarrollar un sistema de traducción automática del habla se requieren dos tipos de recursos en cada una de las lenguas entre las que se realizará la traducción: corpus léxicos y corpus de diálogos; deben ser, además, corpus paralelos, es decir, contener material lingüístico equivalente, y estar alineados, de modo que para cada enunciado pueda identificarse su correspondiente traducción.

Este tipo de recursos se han creado para el español en el marco de iniciativas específicamente dedicadas a la constitución de corpus y a la definición de procedimientos y estándares como LC-STAR (*Lexica and Corpora for Speech-to-Speech Translation Components*) y, naturalmente, en proyectos orientados al desarrollo de sistemas de traducción oral como EUTRANS (*Example-based language TRANslation Systems*) (Pastor *et alii,* 2000), SisHiTra (González *et alii,* 2002) o FAME (*Facilitating Agents in Multicultural Exchange*) (Arranz *et alii,* 2004).

En el caso de LC-STAR (Conejero *et alii,* 2003), por ejemplo, se dispone para el español –en paralelo con el inglés y el catalán– de un léxico de unas 55.000 palabras y de un corpus de 217 diálogos centrados en el dominio del turismo en los que intervienen 77 locutores, además de la traducción española de transcripciones de diálogos del proyecto Verbmobil; según se indica en las publicaciones del proyecto, estos recursos se distribuirán a través de ELDA.

2.8. *Corpus para la recuperación de información*

Otra de las necesidades que recientemente se ha planteado es el acceso a la información almacenada en grandes bases de datos que contienen señales sonoras y visuales, como puedan ser los archivos de emisoras de radio o de televisión. Para desarrollar un sistema que haga posible recuperar los documentos o los datos deseados cuando no se dispone de una transcripción ortográfica, se requieren también corpus que permitan entrenar un reconocedor específicamente dedicado a esta tarea. Por ello, existen ya corpus de noticias en varias lenguas, como los que hemos mencionado en el contexto del reconocimiento del habla, especialmente los conocidos como HUB-4 y 5, sucintamente descritos en la tabla 3.

Se han creado también otros recursos que incluyen el español, como el corpus multimodal bilingüe español-euskera descrito en Bordel *et al.* (2004), con 6 horas de vídeo y audio, y su correspondiente transcripción, extraídas del noticiario *Teleberri* emitido en español por la televisión vasca; la transcripción contiene información sobre cambios de locutor, ruidos o música de fondo. A partir de estos datos se ha elaborado también un léxico con información fonética y morfológica. Por su parte, Transcrigal-DB es un corpus multimodal de noticias en gallego y en español, procedentes del *Telexornal* de la televisión pública gallega, en el que se incluyen intervenciones en español (García *et alii*, 2004).

3. La transcripción y anotación fonéticas de corpus orales

3.1. *La transcripción de los elementos segmentales*

El uso de transcripciones fonéticas en el campo de las tecnologías del habla requiere una codificación a la que puedan acceder los ordenadores; sin embargo, no todos los símbolos del AFI (Alfabeto Fonético Internacional) (International Phonetic Association 1999) –el de uso generalizado en los estudios del habla– tienen su representación en el código ASCII[3], por lo que ya a finales de los años 80 se planteó la necesidad de establecer estándares adecuados a las nuevas necesidades tecnológicas. Entre los alfabetos creados para un uso informático destacan, por la extensión de sus aplicaciones, SAMPA y Worldbet, que se describen a continuación.

---

[3] El sistema de codificación Unicode podrá, probablemente, contribuir a resolver este problema cuando se generalice su uso en los sistemas informáticos (Wells, 2003).

3.1.1. *SAMPA*

SAMPA (*SAM Phonetic Alphabet*), desarrollado inicialmente entre 1987 y 1989 como parte del ya mencionado proyecto SAM, reproduce los símbolos del AFI que coinciden con el alfabeto latino y codifica los restantes con caracteres ASCII de 7 bits (códigos 32 a 127) (Wells, 1997). Con SAMPA se obtiene una transcripción ancha, al igual que con el AFI, puesto que sólo incluyen los símbolos de aquellos segmentos que tienen valor distintivo; los detalles fonéticos deben deducirse del contexto. Por otro lado, como fue ideado para transcribir lenguas europeas, su uso es relativamente limitado. La posterior ampliación de este alfabeto, el llamado X-SAMPA (*Extended SAMPA*) (Wells, 1995a), reproduce todos los signos del AFI, incluyendo los referentes a las características prosódicas de los enunciados, que forman parte de SAMPROSA (*SAM Prosodic Alphabet*) (Wells, 1995b). Por ello, puede ser considerado un estándar aplicable a todos los sistemas lingüísticos, que permite, también, transcripciones más detalladas. Se ha empleado fundamentalmente en proyectos europeos, por ejemplo, en la transcripción del corpus multilingüe EUROM y en las bases de datos diseñadas a partir de los criterios de SpeechDat, descritas en el apartado 2.

SAMPA tiene una adaptación específica para el español (Llisterri y Mariño, 1993) que recoge los símbolos correspondientes a los segmentos con valor distintivo tradicionalmente descritos para la lengua: vocales; africadas sorda y sonora; oclusivas bilabiales, dentales y velares, sordas y sonoras; fricativas sordas labiodental, interdental, dental y velar; nasales bilabial, alveolar y palatal; laterales alveolar y palatal, y vibrantes simple y múltiple. Asimismo, se incluyen los alófonos aproximantes (clasificados como fricativos) y las semivocales, cuyos símbolos, al igual que en el AFI, se usan también para transcribir las semiconsonantes. En la tabla 5, tomada de Wells (1995c), constan los símbolos de SAMPA utilizados en español.

TABLA 5: Símbolos de SAMPA *(SAM Phonetic Alphabet)* para la transcripción del español (Wells 1995c).

SAMPA	Ejemplo	Ejemplo transcrito	SAMPA	Ejemplo	Ejemplo transcrito
p	padre	"paDre	tS	mucho	"mutSo
b	vino	"bino	jj	hielo	"jjelo
t	tomo	"tomo	f	fácil	"faTil
d	donde	"donde	B (= /b/)	cabra	"kaBra
k	casa	"kasa	T	cinco	"Tinko
g	gata	"gata	D (= /d/)	nada	"naDa

TABLA 5: Símbolos de SAMPA *(SAM Phonetic Alphabet)* para la transcripción del español (Wells 1995c). *(Continuación)*

SAMPA	Ejemplo	Ejemplo transcrito
s	sala	"sala
x	mujer	mu"xer
G (= /g/)	luego	"lweGo
m	mismo	"mismo
n	nunca	"nunka
J	año	"aJo
l	lejos	"lexos
L (o como jj)	caballo	ka"baLo
r	puro	"puro
rr	torre	"torre
j	rey pie	"rrej""pje"
w	deuda	"dewDa
i	pico	"piko
e	pero	"pero
a	valle	"baLe
o	toro	"toro
u	duro	"duro

Para transcribir el español dando cuenta de la variación alofónica es necesario recurrir a la extensión X-SAMPA, con la que es posible indicar determinadas características de los sonidos que, de otro modo, quedan omitidas, por ejemplo, la articulación dental de [t̪] y [d̪], y la interdental de [θ̪]. Cabe añadir que, en el marco del proyecto SALA (véase el apartado 2.3.), se desarrolló una adaptación específica de SAMPA para la transcripción de las diversas variedades del español de América (Mariño y Moreno 1998).

3.1.2. *Worldbet*

El planteamiento de Worldbet (Hyeronimus, 1994, 1997) tiene un carácter más ambicioso. Fue ideado para representar todos los sonidos de una amplia variedad de lenguas, por lo que facilita la transcripción de bases de datos multilingües. Se parte del principio de que cualquier sonido que sea espectral y temporalmente distintivo debe ser representado por un símbolo de base; éstos pueden ser modificados con la concatenación de un diacrítico para transcribir los efectos de la coarticulación o del contexto, o las variaciones tonales. Worldbet codifica los símbolos del AFI, además de otros adicionales, no incluidos en dicho alfabeto, útiles para el etiquetado de corpus. En la actualidad, el número total de símbolos es de 299, cada uno representado por dos caracteres ASCII. Un sistema derivado de Worldbet es OGI-bet (Lander, 1997), empleado, por ejemplo, en la transcripción de los corpus distribuidos por el CSLU *(Center for Spoken Language Understanding, Oregon Graduate Institute).*

Una transcripción estrecha en español también se puede obtener mediante los símbolos de Worldbet; en la tabla 6, tomada de Hyeronimus (1994), se recogen los utilizados en este alfabeto para la transcripción fonológica del español peninsular.

TABLA 6: Símbolos de Worldbet para la transcripción del español peninsular (Hyeronimus, 1994).

Worldbet	Ejemplo	Ejemplo transcrito
p	punto	p u n t o
b	baños	b a n~ o s
t	tino	t i n o
d	donde	d o n d e
k	casa	k a s a
g	ganga	g a N g a
V	haba	a V a
f	falda	f a l d a
s	casa	k a s a
z	mismo	m i z m o
T	luces	l u T e s
D	dedo	d e D o
x	jamás	x a m a s
G	lago	l a G o
tS	chato	t S a t o
dZ	un yugo	dZ u G o

Worldbet	Ejemplo	Ejemplo transcrito
m	mano	m a n o
n	nada	n a D a
n~	baño	b a n~ o
N	banco	b a N k o
l	lado	l a D o
L	pollo	p o L o
r(	pero	p e r(o
r	perro	p e r o
j	mayo	m a j o
w	cuento	k w e n t o
i	piso	p i s o
e	mesa	m e s a
a	caso	k a s o
o	modo	m o D o
u	cura	k u r(a

Worldbet y OGIbet se han adaptado al español de México para llevar a cabo la transcripción fonética y fonológica de recursos como los ya mencionados DIME y DIMEx100 (Cuétara, 2004), tomando como base las propuestas de Uraga y Pineda (2002).

3.2. *La transcripción de los elementos suprasegmentales*

Así como para la transcripción de los elementos segmentales existe un procedimiento universalmente aceptado –el Alfabeto Fonético Internacio-

nal– y, como mínimo, dos sistemas ampliamente extendidos –SAMPA, de tradición europea, y Worldbet en el contexto americano–, para la transcripción de los elementos suprasegmentales no se ha llegado a un consenso similar, tal como pone de manifiesto la revisión de Quazza y Garrido (1998), realizada como parte del proyecto MATE, cuyo objetivo era, como se ha indicado, la anotación de corpus. En este trabajo nos centraremos en dos sistemas de representación de la entonación -INTSINT y ToBI- que, aunque no pueda decirse que constituyan todavía un estándar, se han utilizado en varios de los recursos citados en el apartado 2.

3.2.1. *INTSINT*

INTSINT (*INternational Transciption System for INTonation*) es un sistema de anotación prosódica desarrollado principalmente por Daniel Hirst y Albert Di Cristo en el marco del denominado modelo de Aix-en-Provence (Baqué y Estruch, 2003; Di Cristo *et alii*, 2002; Hirst *et alii*, 2000). De modo similar al AFI, INTSINT puede considerarse universal, ya que es posible usarlo como una herramienta para recoger datos sin necesidad de conocer previamente el inventario de patrones tonales de la lengua, contrariamente a lo que sucede, por ejemplo, con ToBI (véase el apartado 3.2.2.).

Para la transcripción de la curva melódica INTSINT utiliza un total de 8 tonos: T (*Top*), M (*Mid*), B (*Bottom*), H (*Higher*), S (*Same*), L (*Lower*), U (*Upstepped*), D (*Downstepped*), de entre los cuales podemos diferenciar los tonos absolutos, interpretados globalmente a partir del rango frecuencial del locutor; y los tonos relativos, interpretados localmente a partir del valor frecuencial del tono anterior y del tono posterior. En el caso de los tonos relativos, además, cabe diferenciar entre los tonos iterativos y los no iterativos. En la tabla 7 se resume la interpretación de cada uno de los tonos.

TABLA 7: Tonos para la representación de la curva melódica mediante INTSINT.

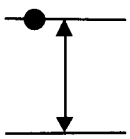
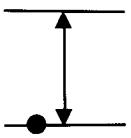
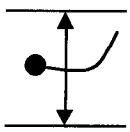
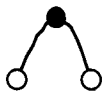
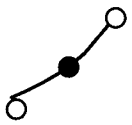
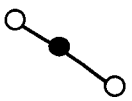
Tonos absolutos		
T (*Top*)	Punto más alto en el rango frecuencial del locutor	
B (*Bottom*)	Punto más bajo en el rango frecuencial del locutor	

TABLA 7: Tonos para la representación de la curva melódica mediante INTSINT. *(Continuación)*

M (*Mid*)	Punto medio en el rango frecuencial del locutor que acostumbra a utilizarse para codificar el inicio de la curva	
Tonos relativos no iterativos		
H (*Higher*)	Punto frecuencial más alto que el anterior y el posterior (pico o máximo local)	
L (*Lower*)	Punto frecuencial más bajo que el anterior y posterior (valle o mínimo local)	
S (*Same*)	Punto frecuencial similar al punto anterior	
Tonos relativos iterativos		
U (*Upstepped*)	Punto frecuencial que se encuentra en una secuencia ascendente o pico con valores frecuenciales muy próximos a los de los puntos anterior y posterior, por lo cual no puede considerarse un máximo local	
D (*Downstepped*)	Punto frecuencial que se encuentra en una secuencia descendente o valle con valores frecuenciales muy próximos a los de los puntos anterior y posterior, por lo cual no puede considerarse un mínimo local	

Estos 8 tonos –3 de los cuales son positivos o ascendentes, 3 son negativos o descendentes y 2 son neutros– pueden representarse mediante símbolos ortográficos o a través de símbolos icónicos, tal como se muestra en la tabla 8.

TABLA 8: Símbolos ortográficos e icónicos del sistema de transcripción INTSINT (Adaptado de Hirst et alii, 2000).

		Positivos	*Neutros*	*Negativos*
Tonos absolutos		T⇑	M⇒	B⇓
Tonos relativos	*No iterativos*	H↑	S→	L↓
	Iterativos	U<		D>

Una de las ventajas de INTSINT desde el punto de vista del tratamiento de corpus reside en que la transcripción de la curva melódica puede realizarse de forma automática usando la herramienta de análisis y síntesis del habla MES (*Motif Environement for Speech*), desarrollada también en el *Laboratoire Parole et Langage* de la Universidad de Provenza (Hirst, 2002). Esta herramienta permite llevar a cabo una serie de operaciones a partir de la señal sonora:

- Visualizar el oscilograma y el espectrograma de un fichero de habla y segmentar y etiquetar la señal.
- Realizar el cálculo y la edición de una curva melódica a partir de 3 métodos distintos de extracción de la frecuencia fundamental (F_0) –método espectral, método de autocorrelación y método temporal AMDF (*Average Magnitude Difference Function*)–, así como combinar los 3 métodos de extracción.
- Estilizar la curva de F_0 con el algoritmo MOMEL (*MOdelisation de MElodie*), que permite describir la curva como una secuencia de puntos de inflexión definiendo una función cuadrática (*quadratic spline function*) que une los puntos mediante parábolas (Hirst y Espesser, 1993), y resintetizar la señal usando PSOLA (*Pitch-Synchronous Overlap and Add*).
- Generar y alinear a la señal de forma automática las etiquetas del sistema de transcripción prosódica INTSINT partiendo de la curva estilizada con MOMEL.
- Generar una curva melódica estilizada a partir de un fichero de etiquetas INTSINT.

Uno de los primeros corpus multilingües que utilizaron el análisis mediante MES y MOMEL y la anotación con INTSINT se constituyó en el marco del proyecto MULTEXT (*Multilingual Text Tools and Corpora*); el resultado de los trabajos llevados a cabo fue la base de datos del mismo nombre, mencionada en el apartado 2.2., así como la creación de un conjunto de herramientas para la anotación automática de corpus prosódicos (Campione *et alii*, 2000; Campione y Véronis, 2001).

3.2.2. *ToBI*

ToBI *(Tones (To) and Break Indices (BI))* surgió como un sistema de anotación prosódica desarrollado entre 1991 y 1994 para el etiquetado de bases de datos orales del inglés, pero pronto inspiró el desarrollo de sistemas de anotación en otras lenguas.

Actualmente ToBI se concibe, más que como un alfabeto para transcribir la entonación, como un instrumento para investigar, que resulta también útil para el etiquetado de extensos corpus de habla. Partiendo de un enfoque fonológico, el contorno entonativo se considera una secuencia de fenómenos tonales discretos, y se representa de forma lineal mediante una cadena autosegmental de tonos (*tones*), mientras que la jerarquía métrica se representa mediante la atribución de valores numéricos a los distintos grados de separación entre unidades prosódicas (índices de disyunción o *break indices*). Tanto para las lenguas tonales como para las entonativas se contemplan únicamente dos niveles tonales, uno alto (*High*, H) y otro bajo (*Low*, L), entendidos como objetivos estáticos que se definen en relación al campo tonal de la secuencia y que contrastan entre sí en el eje paradigmático. Con la implementación fonética se consigue otorgar, mediante interpolación, un contorno melódico apropiado a la secuencia de tonos altos y bajos que compone cada unidad melódica (Beckman *et alii*, 2005; Hualde, 2003; Silverman *et alii*, 1992).

Existen una serie de convenciones para la anotación de corpus (Beckman y Ayers, 1997) y, por otra parte, las etiquetas propias de ToBI pueden emplearse en sistemas diseñados para analizar y etiquetar archivos de sonido (como Praat o PitchWorks), para crear y analizar bases de datos (por ejemplo, EMU) y en cualquier otra herramienta que pueda manejar archivos *xlabel*.

La adaptación al español de ToBI, conocida como Sp-ToBI (Beckman *et alii*, 2002; Sosa, 2003), surgió en 1999 con el propósito de resultar adecuada para todas las variedades de la lengua. La tabla 9 resume las principales convenciones en lo que se refiere al nivel de índices de disyunción y al nivel tonal; ToBI contempla también otros niveles de anotación como el ortográfico o de palabras, el silábico, para el que puede emplearse un sistema de transcripción cono SAMPA (véase el apartado 3.1.1), un nivel misceláneo para los fenómenos propios del habla espontánea que alteran la fluidez o para la incorporación de análisis alternativos y, finalmente, un nivel de código que contiene, si se conoce, información sobre el dialecto y el sociolecto del hablante y sobre los cambios de código entonativo.

TABLA 9: Índices de disyunción y tonos para la anotación prosódica mediante ToBI.

Nivel de índices de disyunción	
0	Indica el grado máximo de cohesión. No se percibe separación entre palabras
1	Juntura «ordinaria» entre palabras
2, 3	Se percibe separación y existen indicios claros. Serían unidades candidatas para este tipo de disyunción la frase intermedia el grupo clítico y el grupo tónico (acentual)
4	Se percibe separación y existen indicios claros que permiten identificar la juntura entre grupos melódicos

TABLA 9: Índices de disyunción y tonos para la anotación prosódica mediante ToBI *(continuación)*.

Nivel de tonos	
Acentos tonales	
L*+H	Valle (L) asociado a la sílaba acentuada (*) seguido de (+) un pico (H) Movimiento ascendente iniciado en la sílaba acentuada.
L+H*	Valle (L) seguido de (+) un pico (H) asociado a la sílaba acentuada (*) Movimiento ascendente culminado en la sílaba acentuada.
H+L*	Pico (H) seguido de (+) un valle (L) asociado a la sílaba acentuada (*) Movimiento descendente culminado en la sílaba acentuada.
Tonos de frontera	
L%	Descenso hacia una frecuencia más grave tras tonos como L+H* o mantenimiento de una frecuencia grave tras H+L*
H%	Ascenso hacia una frecuencia más aguda después de cualquier acento tonal.

Además de las mencionadas en la tabla 9, existen otras etiquetas provisionales, cuyo uso se recomienda cuando exista incertidumbre y hasta que sea posible revisar el análisis del dialecto estudiado. De hecho, Sp-ToBI se encuentra todavía en fase de desarrollo y se plantean diferencias importantes entre enfoques a la hora de caracterizar el nivel tonal. Tal es el caso, entre otros, de la adaptación de ToBI al español de Argentina realizada por Gurlekian *et alii* (2001b) para el etiquetado de un corpus prosódico orientado al desarrollo de sistemas de conversión de texto en habla (véase el apartado 2.2); estos autores utilizan ocho tonos distintos y codifican también otros factores, como un índice de percepción de los acentos y la distancia en sílabas entre ellos. Quedan pues por estudiar un buen número de fenómenos entonativos que podrían afectar al inventario de tonos, y se precisa más información sobre los distintos dialectos y sobre las unidades relevantes en la agrupación prosódica.

3.3. *Las herramientas de etiquetado fonético*

Para finalizar este trabajo nos referiremos brevemente a las herramientas para el etiquetado fonético de corpus orales. Entre los numerosos sistemas disponibles, nos centraremos en dos programas –Praat y WaveSurfer– que reúnen unos requisitos que nos parecen esenciales: por una parte, son de dominio público; por otra, se trata de programas que funcionan sobre diversos sistemas operativos como Windows, MacOS o Linux. Vale la pena también recordar que se actualizan constantemente, en ocasiones a partir de las

sugerencias de los usuarios, y ofrecen una documentación adecuada. Para obtener información sobre otras herramientas, remitimos al lector a revisiones como las de Cosi (2002) o Dybkjaer *et alii* (2001)[4].

Praat, un programa creado por Paul Boersma y David Weenink, del Instituto de Fonética de la Universidad de Ámsterdam, permite realizar todo tipo de análisis acústicos, así como el etiquetado fonético de corpus orales (Boersma, 2001). Una de sus ventajas es que muchos investigadores han desarrollado programas complementarios (*scripts*) que permiten ampliar sus funciones y que pueden encontrarse fácilmente en Internet.

El etiquetado se realiza en un fichero de etiquetas independiente del documento que contiene la señal sonora, sincronizado con la representación oscilográfica y espectrográfica del enunciado. Es posible emplear diversos niveles de etiquetado –por ejemplo un nivel fonético, otro fonológico, un tercer nivel con la transcripción ortográfica de las palabras, etc.–, así como añadir o eliminar niveles. Puede también emplearse al AFI si se ha instalado previamente la fuente SIL Doulos IPA 1989. En la figura 1 se muestra un ejemplo de etiquetado de un fragmento mediante el programa Praat.

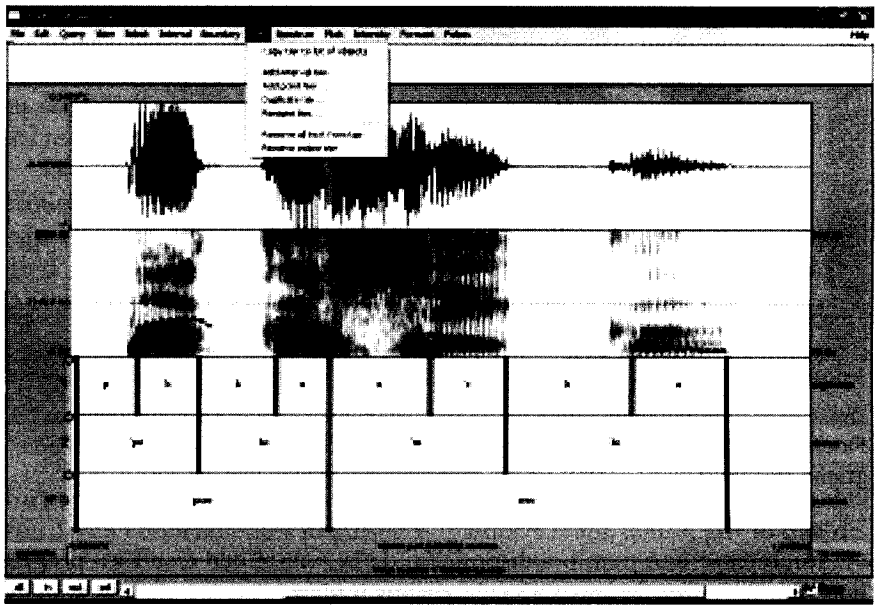


FIGURA 1: Etiquetado en tres niveles (fonemas, sílabas y palabras) de la secuencia *poco seco* mediante el programa Praat.

[4] Destacaríamos, sin embargo, Transcriber, una herramienta de transcripción, segmentación y etiquetado de corpus orales, también multiplataforma y de dominio público desarrollada por Claude Barras y Edouard Geoffrois. (Barras et *alii*, 2001).

WaveSurfer, creado por Kåre Sjölander y Jonas Beskow, del Centro de Tecnología del Habla del KTH (Real Instituto de Tecnología) de Estocolmo, es también una herramienta orientada al análisis acústico y al etiquetado fonético de corpus orales (Sjölander y Beskow, 2000).

El programa ofrece varios formatos de transcripción, y ésta se realiza a partir de la representación espectrográfica y oscilográfica de la señal sonora. El conjunto de etiquetas, con la información temporal necesaria para sincronizarlas con la grabación, se almacena en un fichero de texto que puede editarse para realizar modificaciones. Es también posible cambiar el tipo de fuente y el formato de la transcripción. En la figura 2 se observa el mismo fragmento de la figura 1, etiquetado aquí empleando WaveSurfer.

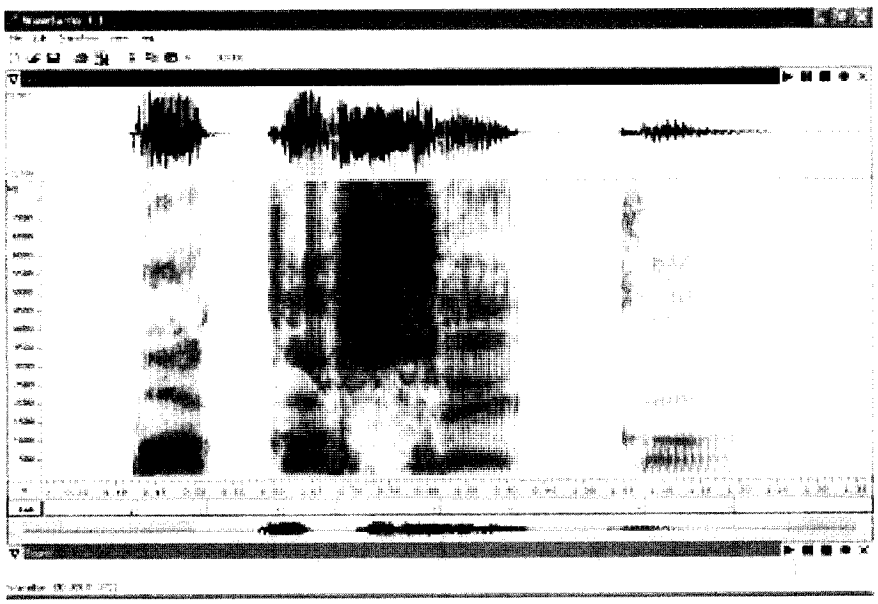


FIGURA 2: Etiquetado fonético segmental de la secuencia *poco seco* mediante el programa WaveSurfer.

También se han puesto a disposición de la comunidad científica programas que permiten la anotación de corpus multimodales, tratando simultáneamente la señal sonora y la visual; esto es especialmente útil en el contexto de las bases de datos que se contemplan en el desarrollo de sistemas de diálogo, tal como se hacía patente en el apartado 2.6. Entre las herramientas de dominio público podemos destacar Anvil (Kipp, 2001), que permite emplear ficheros procedentes de Praat, ELAN (*EUDICO Linguistic Annotator*, Instituto Max Planck de Psicolingüística) o Transana (Universidad de Wisconsin).

4. Conclusiones

Este trabajo ha pretendido mostrar la especificidad de un tipo de recurso lingüístico, los corpus orales empleados en el desarrollo de tecnologías del habla, presentando sucintamente los materiales que hemos logrado documentar, los sistemas para el etiquetado fonético segmental y suprasegmental y las herramientas de dominio público que permiten llevar a cabo una anotación. Debe señalarse que SAMPA, Worldbet, INSINT y ToBI se usan también en los estudios fonéticos llevados a cabo desde una perspectiva lingüística, y que programas como Praat y WaveSurfer son habituales en los trabajos sobre análisis acústico del habla; no es ya tan frecuente el uso sistemático de corpus como los aquí mencionados, pero indudablemente son recursos que pueden aprovecharse cuando se dispone de ellos.

Esto nos lleva, precisamente, a constatar que si bien se cuenta con herramientas, es difícil, en muchas ocasiones, acceder a los datos. Se ha señalado repetidamente que la adquisición de un corpus es una operación sumamente costosa en recursos humanos –y, por tanto, económicos–, por lo que no deben resultar extraños los precios de venta que aparecen en los catálogos de ELDA y del LDC. Existen en español bases de datos orales creadas con financiación pública en el marco de diversos proyectos y soluciones para su difusión[5] pero, aun así, parece que no siempre es sencillo ni localizar estos recursos ni conseguirlos.

Es también un lugar común referirse a la reutilización de los corpus existentes. Por el momento, hay motivos para pensar que, en lo que se refiere al español, no se ha logrado avanzar mucho en esta línea. Es posible que ello se deba a una escasez de información centralizada, a la falta de continuidad de las iniciativas que han intentado reunirla o que han tenido como objetivo estudiar la posibilidad de difundir recursos en formatos compatibles y, principalmente, a los propios mecanismos de planificación científica y tecnológica. que no siempre favorecen, en la práctica, la estandarización ni los medios necesarios para poder intercambiar datos y herramientas (Llisterri, 2004).

Finalmente, cabe recordar que muchas de las etapas del desarrollo de un corpus enfocado a las aplicaciones en el campo de las tecnologías del habla requieren conocimientos lingüísticos (Llisterri *et alii*, 2003), por lo que debería ser una práctica habitual incorporar a los equipos implicados en el proceso de constitución de recursos a especialistas que pudieran asegurar una buena calidad también en este ámbito.

---

[5] Por ejemplo, ELDA distribuye Albayzín (realizado enteramente con financiación pública española) al precio de 120 euros para los centros de investigación españoles, de 2000 euros para los centros de investigación extranjeros y de 12000 euros para empresas extranjeras, siempre en el caso de que no sean socios de ELDA. Es, sin embargo, el único caso que conocemos para el que rigen estas condiciones.

BIBLIOGRAFÍA[6]

Actas de las I Jornadas en Tecnologías del Habla, Sevilla, del 6 al 10 de noviembre de 2000, Sevilla, Universidad de Sevilla.

ALCÁCER, N., M.J CASTRO, I. GALIANO, R. GRANELL, S. GRAU y D. GRIOL (2004): «Adquisición de un corpus de diálogo: DIHANA», en: E. Sanchis (ed.), págs. 131-136.
http://www.iti.upv.es/~prhlt/PAPERS/papers01-05/2004/Alcacer04a.pdf

ARRANZ, V., E. COMELLES y D. FARWELL (2004): «Sistema de traducción oral de ayuda al intercambio multicultural», en: E. Sanchis (ed.), págs. 189-194.
http://isl.ira.uka.de/fame/publications/FAME-A-WP9-023.pdf

BAQUÉ, L. y M. ESTRUCH (2003): «Modelo de Aix-en-Provence», en: P. Prieto (ed.), págs. 123-154.
http://seneca.uab.es/lorraine_baque/Publications/ModeloAix-en-ProvenceV3.pdf

BARRAS, C., E. GEOFFROIS, Z. WU y M. LIBERMAN (2001): «Transcriber: development and use of a tool for assisting speech corpora production», *Speech Communication*, 33, 1-2, págs. 5-22. .
http://www.etca.fr/CTA/gip/Projets/Transcriber/articles/Transcriber-Speech-Comm2000.ps

BATTANER, E., J. GIL, V. MARRERO, J. LLISTERRI, C. CARBÓ, M.J. MACHUCA, C. DE LA MOTA y A. RÍOS (2003): «VILE: Estudio acústico de la variación inter e intra locutor en español», en: *SEAF 2003, Actas del II Congreso de la Sociedad Española de Acústica Forense*, Barcelona, 10 y 11 de abril de 2003, Barcelona, SEAF, Sociedad Española de Acústica Forense, págs. 59-70.
http://liceu.uab.es/~joaquim/phonetics/VILE/VILE_SEAF03.pdf

BECKMAN, M.E. y G.M. AYERS (1997): *Guidelines for ToBI Labelling*, Version 3, March 1997, Department of Linguistics, Ohio State University.
http://ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf

BECKMAN, M.E., J. HIRSCHBERG y S. SHATTUCK-HUFNAGEL (2005): «The original ToBI system and the evolution of the ToBI framework», en: S.-A. Jun (ed.), *Prosodic Typology. The Phonology of Intonation and Phrasing*, Oxford, Oxford University Press.
http://www.ling.ohio-state.edu/~tobi/JunBook/BeckHirschShattuckToBI.pdf

BECKMAN, M.E., M. DÍAZ CAMPOS, J.T. MCGORY y T.A. MORGAN (2002): «Intonation across Spanish, in the Tones and Break Indices framework», *Probus*, 14, 1, págs. 9-36.
http://www.ling.ohio-state.edu/~mbeckman/Sp_ToBI/Sp_ToBI_Jul29.pdf

BOERSMA, P. (2001): «Praat, a system for doing phonetics by computer», *Glot International*, 5, 9-10, págs. 341-345.

BONAFONTE, A., P. AIBAR, N. CASTELL, E. LLEIDA, J.B. MARIÑO, E. SANCHIS y M.I. TORRES

[6] Las direcciones de Internet que se recogen en este apartado se han verificado por última vez en marzo de 2005.

(2000): «Desarrollo de un Sistema de Diálogo Oral en Dominios Restringidos», en: *Actas de las I Jornadas en Tecnologías del Habla.*
http://gps-tsc.upc.es/veu/basurde/download/Bon00a_sevilla.pdf

BORDEL, G., A. EZEIZA, K. LÓPEZ DE IPIÑA, M. MÉNDEZ, M. PEÑAGARIKANO, T. RICO, C. TOVAR y ZULUETA (2004): «Development of Resources for a Bilingual Automatic Index System of Broadcast News in Basque & Spanish», en: *LREC 2004*, págs. 881-884.

CAMPIONE E. y J. VÉRONIS (1998): «A Multilingual Prosodic Database», en: *ICSLP 1998*, págs. 3163-3166.
http://www.up.univ-mrs.fr/~veronis/pdf/1998icslp-database.pdf

CAMPIONE E. y J. VÉRONIS (2001): «Etiquetage prosodique semi-automatique des corpus oraux», en: *TALN'2001, Actes de la Conférence «Traitement Automatique des Langues»*, 2 - 5 juillet 2001, Tours, France, págs. 123-132.
http://www.up.univ-mrs.fr/~veronis/pdf/2001-taln.pdf

CAMPIONE, E., D. HIRST y J. VÉRONIS (2000): «Automatic stylisation and symbolic coding of F0: Implementations of the INTSINT model», en: A. Botinis (ed.), *Intonation: Analysis, Modelling and Technology*, Dordrecht, Kluwer Academic Publishers, págs. 185-208.
http://www.up.univ-mrs.fr/~veronis/pdf/2000Campione.pdf

CASACUBERTA, F., R. GARCÍA, J. LLISTERRI, C. NADEU, J.M. PARDO y A. RUBIO (1992): «Desarrollo de corpus para investigación en tecnologías del habla (Albayzín)», *Procesamiento del Lenguaje Natural*, 12, págs. 35-42.
http://www.sepln.org/revistaSEPLN/revista/12/12-Pag35.pdf
http://liceu.uab.es/~joaquim/publicacions/Casacuberta_et_al_92.pdf

CHAN, D., A. FOURCIN, D. GIBBON, B. GRANSTRÖM, M. HUCKVALE, G. KOKKINAKIS, K. KVALE, L. LAMEL, B. LINDBERG, A. MORENO, J. MOUROPOULOS, F. SENIA, I TRANCOSO, C. VELD y J. ZEILIGER (1995): «EUROM- A Spoken Language Resource for the EU», en: *EUROSPEECH 1995*, págs. 867-870.
http://www.phon.ucl.ac.uk/resource/eurom1/eurospeech95eurom.pdf

CIERI, C., J.P. CAMPBELL, H. NAKASONE, D. MILLER y K. WALKER (2004): «The Mixer corpus of multilingual, multichannel speaker recognition data», en: *LREC 2004*, págs. 627-630.

CONEJERO, D., J. GIMÉNEZ, V. ARRANZ, A. BONAFONTE, N. PASCUAL, N. CASTELL y A. MORENO (2003): «Lexica and corpora for speech-to-speech translation: A trilingual approach», en: *Eurospeech 2003 - Interspeech 2003, Proceedings of the 8h European Conference on Speech Communication and Technology*, 1 - 4 September, 2003, Geneva, Switzerland, págs. 1593-1596.
http://gps-tsc.upc.es/veu/research/pubs/download/Con_lex_03.pdf http://www.lc-star.com/Con_lex_03.pdf

COSI, P. (2002): «Metodologie e sistemi per l'annotazione linguistica», *Quaderni dell'Istituto di Fonetica e Dialettologia*, 4.
http://www.csrf.pd.cnr.it/Papers/quaderni2002.zip

CUÉTARA, J.O. (2004) *Fonética de la ciudad de México. Aportaciones desde las tecnologías del habla*, Tesis de Maestría en Lingüística Hispánica, Posgrado en Lingüística, Universidad Nacional Autónoma de México.
http://www.filos.unam.mx/LICENCIATURA/VOX/JAVIER/TesisMLH_Cuetara/Indice.htm [username: javier; password: cuetara]

Di Cristo, A., D. Hirst, N. Boudouresques y M. Louis (2002): «Écrire l'intonation: le système INTSINT, fondements théoriques et illustrations», *Revue PArole*, 22-23-24, págs. 175-212.

Draxler, C., H. van den Heuvel y H. Tropf (1998): «SpeechDat Experiences in Creating Large Multilingual Speech Databases for Teleservices», en: *LREC 1998, Proceedings of the First International Conference on Language Resources and Evaluation*, 28 - 30 May 1998. Granada, Spain, Paris, ELRA, European Language Resources Association, págs. 361-366.

http://lands.let.kun.nl/literature/heuvel.1998.1.ps

Dybkjaer, L., S. Berman, M. Kipp, M. Wagener, V. Pirrelli, N. Reithinger y C. Soria (2001): *Survey of Existing Tools, Standards and User Needs for Annotation of Natural Interaction and Multimodal Data*, ISLE Natural Interactivity and Multimodality Working Group, D11.1., January 2001.

http://isle.nis.sdu.dk/reports/wp11/

Esquerra, I., A. Bonafonte, F. Vallverdú y A. Febrer (1998): «A bilingual Spanish-Catalan database of units for concatenative synthesis», en: *Workshop on Language Resources for European Minority Languages*, May 27, 1998. Granada, Spain, págs. 39-42.

http://193.2.100.60/SALTMIL/files/ignasi1.zip

Esteve, J., D. Tapias y J.C. Torrecilla (1994): «La base de datos VESTEL», *Comunicaciones de Telefónica I+D*, 5, 2, págs. 44-54.

http://www.tid.es/presencia/publicaciones/comsid/esp/articulos/vol52/artic3/3.html

Eurospeech 1995, Proceedings of the 4th European Conference on Speech Communication and Technology, 18 - 21 September, 1995, Madrid, Spain.

García, C., J. Diéguez, L. Docío y A. Cardenal (2004): «Transcrigal: A bilingual system for automatic indexing of broadcast news», en: *LREC 2004*, págs. 2061-2064.

Gibbon, D., R. Moore y R. Winski (eds.) (1997): *Handbook on Standards and Resources for Spoken Language Systems*, Berlin, Mouton De Gruyter.

González, J., J.R. Navarro, F. Nevado, M. Pastor, F. Casacuberta, E. Vidal, F. Fabregat, J.M. De Val, L. Arenas, F. Pla y J. Tomás (2002): «SisHiTra: Sistemas de traducción catalán-castellano y castellano-catalán con entrade de texto y voz», en: A. Rubio (ed.) *Actas de las II Jornadas en Tecnologías del Habla*, Granada, del 16 al 18 de diciembre de 2002, Granada, Universidad de Granada, Departamento de Electrónica y Tecnología de Computadores.

http://www.rthabla.org/TECHABLA02/articulos/7.pdf

Gurlekian, J., H. Rodríguez, L. Colantoni y H. Torres (2001b): «Development of a prosodic database for an Argentine Spanish text to speech system», en: *Proceedings of the IRCS Workshop on Linguistic Databases*, págs. 99-104.

http://www.ldc.upenn.edu/annotation/database/papers/Gurlekian_etal/33.3.pdf

Gurlekian, J., L. Colantoni, H. Torres, A. Rincón, A. Moreno y J.B. Mariño (2001a): «Database for an automatic speech recognition system for Argentine Spanish», en: *Proceedings of the IRCS Workshop on Linguistic Databases*, págs. 92-98.

http://www.ldc.upenn.edu/annotation/database/papers/Gurlekian_etal/36.2.gurlekian.pdf

Hennebert, J., H. Melin, D. Petrovska y S. Genoud (2000): «POLYCOST: A telephone-speech database for speaker recognition», *Speech Communication*, 31, 2-3, págs. 265-270.

HEUVEL, H. VAN DEN, P. HALL, H. HÖGE, A. MORENO, A. RINCÓN y F. SENIA (2004): «SALA II across the finish line : a large collection of mobile telephone speech databases from North & Latin America completed», en: *LREC 2004*, págs. 97-100.
http://gps-tsc.upc.es/veu/research/pubs/download/Heu_SAL_04.pdf
http://lands.let.kun.nl/literature/heuvel.2004.1.pdf

HIERONYMUS, J.L. (1994): *ASCII phonetic symbols for the world's languages: Worldbet*, AT&T Bell Laboratories, Technical Report.
http://www.ling.gu.se/~jimh/courses/ipa.ps

HIERONYMUS, J.L. (1997) *Worldbet Phonetic Symbols for Multilanguage Speech Recognition and Synthesis*, AT&T Bell Laboratories, Technical Report.
http://www.ling.gu.se/~jimh/courses/ipa.recog.unicode.ps

HIRST, D.J. (2002): «Automatic analysis of prosody for multilingual speech corpora», en: E. Keller, C. Bailly, A. Monaghan, J. Terken y M. Huckvale (eds.), *Improvements in Speech Synthesis. Cost 258: The Naturalness of Synthetic Speech*, Chichester, John Wiley & Sons, págs. 320-327.
http://www.lpl.univ-aix.fr/~hirst/articles/2001%20Hirst.pdf

HIRST, D.J. y R. ESPESSER (1993): «Automatic modelling of fundamental frequency using a quadratic spline function», *Travaux de l'Institut de Phonétique d'Aix*, 15, págs. 71-85.
http://www.lpl.univ-aix.fr/~hirst/articles/1993%20Hirst&Espesser.pdf

HIRST, D.J., A. DI CRISTO y R. ESPESSER (2000): «Levels of representation and levels of analysis for the description of intonation systems», en: M. Horne (ed.) *Prosody: Theory and Experiment. Studies presented to Gösta Bruce*, Dordrecht, Kluwer Academic Publishers, págs. 51-88.
http://www.lpl.univ-aix.fr/~hirst/articles/2000%20Hirst&al.pdf

HOZJAN, V., C. KACIC, A. MORENO, A. BONAFONTE y A. NOGUEIRAS (2002): «Interface Databases: Design and Collection of a Multilingual Emotional Speech Database», en: *LREC 2002*, págs. 2024-2028.
http://gps-tsc.upc.es/veu/research/pubs/download/hoz_int_02.pdf

HUALDE, J.I. (2003): «El modelo métrico y autosegmental», en: P. Prieto (ed.), págs. 155-184.

ICSLP 1992, Proceedings of the 2nd International Conference on Spoken Language Processing, 12 - 16 October, 1992, Banff, Alberta, Canada, Edmonton, The University of Alberta.

ICSLP 1998, Proceedings of the 5th International Conference on Spoken Language Processing, 30 November - 4 December, 1998, Sydney, Australia, Australian Speech Technology Association.

INTERNATIONAL PHONETIC ASSOCIATION (1999): *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, Cambridge, Cambridge University Press.

IRIONDO, I., R. GUAUS, A. RODRÍGUEZ, P. LÁZARO, N. MONTOYA, J.M. BLANCO, D. BERNADAS, J.M. OLIVER, D. TENA Y LONGHI (2000): «Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques», en: *Proceedings of the ISCA Workshop on Speech and Emotion: A conceptual framework for research*, 5-7 September 2000, Belfast, Northern Ireland, págs. 161-166.
http://web.salleurl.edu/~iriondo/publicacions/iriondo.pdf

ISKRA, D., B.- GROSSKOPF, K. MARASEK, H. VAN DEN HEUVEL, F. DIEHL y A. KIESSLING (2002): «SPEECON Speech Databases for Consumer Devices: Database Specification and Validation», en: *LREC 2002*, págs. 329-333.
http://gps-tsc.upc.es/veu/research/pubs/download/Die_Spe_02.pdf
KIPP, M. (2001): «Anvil - A generic tool for multimodal dialogue», en: *EUROSPEECH 2001 - INTERSPEECH 2001, Proceedings of the 7th European Conference on Speech Communication and Technology*, 3 - 7 September, 2001, Aalborg, Denmark, págs. 1367-1370.
http://www.dfki.de/~kipp/public_archive/kipp2001-eurospeech.pdf
KLEIN, M., N.O. BERNSEN, S. DAVIES, L. DYBKJAER, J.M. GARRIDO, H. KASCH, A. MENGEL, V. PIRRELLI, M. POESIO, S. QUAZZA y C. SORIA (1998): *Supported Coding Schemes*, MATE Deliverable D1.1., LE Telematics Project LE4–8370, July 1998.
http://mate.nis.sdu.dk/about/D1.1/
LAMEL, L.F., G. ADDA, M. DDA-DECKER, C. CORREDOR-ARDOY, J.J. GANGOLF y J.L. GAUVAIN (1998): «A Multilingual Corpus for Language Identification», en: *LREC 1998, Proceedings of the First International Conference on Language Resources and Evaluation*, 28 - 30 May 1998, Granada, Spain, Paris, ELRA, European Language Resources Association, págs. 1115-1122.
ftp://tlp.limsi.fr/public/lrec98ideal.ps.Z
LANDER, T. (1997): *The CSLU Labeling Guide*, Center for Spoken Language Understanding, Oregon Graduate Institute.
http://cslu.cse.ogi.edu/corpora/docs/labeling.pdf
LANDER, T.L., R.A. COLE, B. OSHIKA y M. NOEL (1995): «The OGI 22 Language Telephone Speech Corpus», in *EUROSPEECH 1995*, págs. 817-820.
LLISTERRI, J. (1996): *Preliminary Recommendations on Spoken Texts*, EAGLES Document EAG-TCWG-STP/P, May 1996.
http://www.ilc.cnr.it/EAGLES96/spokentx/spokentx.html
LLISTERRI, J. (2003): «Las tecnologías del habla para el español», en: *Seminario «Ciencia, Tecnología y Lengua Española: La terminología científica en español»*, Madrid, 11 y 12 de diciembre de 2003, Fundación Española para la Ciencia y la Tecnología.
http://liceu.uab.es/~joaquim/publicacions/TecnolHablaEsp_FECyT03.pdf
LLISTERRI, J. (2004): «Las tecnologías lingüísticas en España», en: *El español en el mundo. Anuario del Instituto Cervantes 2004*, Madrid, Instituto Cervantes, Círculo de Lectores, Plaza & Janés, págs. 229-251.
LLISTERRI, J. y J.B. MARIÑO (1993): *Spanish adaptation of SAMPA and automatic phonetic transcription*, SAM-A/UPC/001/v1, 20th April 1993, ESPRIT PROJECT 6819 (SAM-A Speech Technology Assessment in Multilingual Applications).
http://liceu.uab.es/~joaquim/publicacions/SAMPA_Spanish_93.pdf
LLISTERRI, J., C. CARBÓ, M.J. MACHUCA, C. DE LA MOTA, M. RIERA y A. RÍOS (2004a): «La conversión de texto en habla: aspectos lingüísticos», en M.A. Martí y J. Llisterri (eds.), págs. 145-186.
http://liceu.uab.es/publicacions/Linguistica_CTH_FDS02.pdf
LLISTERRI, J., C. CARBÓ, M.J. MACHUCA, C. DE LA MOTA, M. RIERA y A. RÍOS (2003): «El papel de la lingüística en el desarrollo de las tecnologías del habla», en: M. Casas (dir.) y C. Varo (ed.) *VII Jornadas de Lingüística*, Cádiz, Servicio de Publicaciones de la Universidad de Cádiz, págs. 137-191.
http://liceu.uab.es/publicacions/Linguistica_TH_Cadiz02.pdf

LLISTERRI, J., M.J. MACHUCA, N. MADRIGAL, F. MANCINI, P. MASSIMINO, C. DE LA MOTA, M. RIERA Y A. RÍOS (2004b): «Aspectos lingüísticos en el diseño de un conversor de texto en habla en castellano y en catalán: El sistema LoquendoTTS®», en: *Actas del 6º Congreso de Lingüística General*, Santiago de Compostela, del 3 al 7 de mayo de 2004, Área de Lingüística Xeral, Universidade de Santiago de Compostela (en prensa).

http://liceu.uab.es/~joaquim/publicacions/CLG04_Loquendo.pdf

LREC 2000, Proceedings of the Second International Conference on Language Resources and Evaluation, 31 May - 2 June, 2000, Athens, Greece, Paris, European Language Resources Association.

LREC 2002, Proceedings of the Third International Conference on Language Resources and Evaluation, 27 May - 2 June, 2002, Las Palmas de Gran Canaria, Spain, Paris, European Language Resources Association

LREC 2004, Proceeedings of the 4th International Conference on Language Resources and Evaluation, 26-28 May, 2004, Lisbon, Portugal, Paris, ELRA, European Language Resources Association.

MARIÑO, J.B. y A. MORENO (1998): «Spanish Dialects: Phonetic Transcription», en: *ICSLP 1998*.

http://www.sala2.org/0327_v2.ps

MARIÑO, J.B. y C. NADEU (2004): «La representación de la voz para el reconocimiento del habla», en: M.A. Martí y J. Llisterri (eds.), págs. 187-224.

MARTÍ, M.A. y J. LLISTERRI (eds.) (2004) *Tecnologías del texto y del habla*, Barcelona, Edicions de la Universitat de Barcelona, Fundación Duques de Soria.

MARTÍNEZ, C.D., E. SANCHIS, F. GARCÍA y P. AIBAR (2002): «A labelling proposal to annotate dialogues», en: *LREC 2002*, págs. 1577-1582.

MONTERO, J.M., J. GUTIÉRREZ, S. PALAZUELOS, E. ENRÍQUEZ, S. AGUILERA y J.M. PARDO (1998): «Emotional speech synthesis: From speech database to TTS», en: *ICSLP 1998*.

http://www-gth.die.upm.es/~juancho/conferences/paper_icslp98_a4_2.pdf

MORENO, A., D. POCH, A. BONAFONTE, E. LLEIDA, J. LLISTERRI, J.B. MARIÑO y C. NADEU (1993): «ALBAYZÍN Speech Database: Design of the Phonetic Corpus», en: *EUROSPEECH 1993, Proceedings of the 3rd European Conference on Speech Communication and Technology*, 21 - 23 September, 1993, Berlin, Germany, págs. 175-178.

http://liceu.uab.es/~joaquim/publicacions/Moreno_et_al_93.pdf

MORENO, A., B. LINDBERG, C. DRAXLER, G. RICHARD, K. CHOUKRI, S. EULER Y J. ALLEN (2000): «SPEECHDAT-CAR. A Large Speech Database for Automotive Environments», en: *LREC 2000*, págs. 895-900.

http://gps-tsc.upc.es/veu/research/pubs/download/Mor00c.pdf http://www.speechdat.org/SP-CAR/CONFEREN/LREC2000.PDF

MORENO, A., F. SENIA y A. RINCÓN (2002): *The complete SALA II project specifications*, Version 1.6., SALA II Technical Report, November 29, 2002.

http://www.sala2.org/SALA%20II_Specifications_v1.6.pdf

MORENO, A., R. COMEYNE, K. HASLAM, H. VAN DEN HEUVEL, H. HÖGE, S. HORBACH y G. MICCA (2000): «SALA: SpeechDat across Latin America. Results of the First Phase», en: *LREC 2000*, págs. 877-882.

http://gps-tsc.upc.es/veu/research/pubs/download/Mor00b.pdf

MUTHUSAMY, Y.K., R.A. COLE y B.T. OSHIKA (1992): «The OGI multi-language telephone speech corpus», en: *ICSLP 1992*, págs. 895-898.

ORTEGA, A., F. SUKNO, E. LLEIDA, A. FRANGI, A. MIGUEL, L. BUERA y E. ZACUR, E. (2004): «AV@CAR: A Spanish Multichannel Multimodal Corpus for In-Vehicle Automatic Audio-Visual Speech Recognition», en: *LREC 2004*, págs. 763-766.

http://www.visionrt.com/Research/lrec04def2.pdf.pdf http://diec.unizar.es/intranet/articulos/uploads/lrec04def2.pdf.pdf

ORTEGA, J., J. GONZÁLEZ y V. MARRERO (2000): «AHUMADA: A large corpus in Spanish for speaker characterization and identification», *Speech Communication*, 31, 2-3, págs. 255-264.

http://www.atvs.diac.upm.es/publicaciones/docs/Ort00a.pdf

PASTOR, M., E. SANCHIS, F. CASACUBERTA y E. VIDAL (2000): «Eutrans: prototipo de traducción automática de voz a voz», en: *Actas de las I Jornadas en Tecnologías del Habla.*

http://www.iti.upv.es/~prhlt/PAPERS/ltu/2000/Pastor00a.pdf

PINEDA, L.A., A. MASSÉ, M. MEZA, E. SALAS, E. SCHWARZ, E. URAGA y L. VILLASEÑOR (2002): «The DIME project», en: C. Coello, A. de Albornoz, L.E. Sucar y O. Cairó (eds.) *MICAI 2002, Proceedings of the Second Mexican International Conference on Artificial Intelligence*, April 22-26, 2002, Mérida, Yucatán, México, Berlin: Springer, págs. 166-175.

PINEDA, L.A., L. VILLASEÑOR, J. CUÉTARA, H. CASTELLANOS y I. LÓPEZ (2004): «DIMEx100: A new phonetic and speech corpus for Mexican Spanish», en: C. Lemaitre, C.A. Reyes y J.A González (eds.) *Iberamia 2004, Proceedings of the 9th Iberoamerican Conference on Artificial Intelligence.* November 22-26, 2004, Puebla, México, Berlin, Springer, págs. 974-983.

PRIETO, P. (ed.) (2003): *Teorías de la entonación*, Barcelona, Ariel.

Proceedings of the IRCS Workshop on Linguistic Databases, 11-13 December 2001, University of Pennsylvania, Philadelphia, PA, USA.

QUAZZA, S. y J.M GARRIDO (1998): «Prosody», en: M. Klein *et alii.*

http://liceu.uab.es/publicacions/MATED1.1.6Prosody/D11_6_Prosody.html

RENATO, A.C. y J.A. ÁLVAREZ (2004): «Corpora of Latin American Spanish for research in prosody and synthesis», en: *SSW5 2004, Proceedings of the 5th ISCA Tutorial and Research Workshop on Speech Synthesis*, 14 -16 June, 2004, Oakland, Pittsburgh, PA, USA, págs. 221-222.

http://www.ssw5.org/notes/2028.pdf

RODRÍGUEZ, A., P. LÁZARO, N. MONTOYA, J.M. BLANCO, D. BERNADAS, J.M. OLIVER y L. LONGHI (1999): «Modelización acústica de la expresión emocional en español», *Procesamiento del Lenguaje Natural*, 25, págs. 159-166.

http://www.sepln.org/revistaSEPLN/revista/25/25-Pag159.pdf

RODRÍGUEZ, L., C. GARCÍA y J.L. ALBA (2003): «On combining classifiers for speaker authentication», *Pattern Recognition Journal*, 36, págs. 347-359.

http://www.gts.tsc.uvigo.es/image_speech/Journal_papers/PR2003.pdf

SANCHIS, E. (ed.) (2004): *Actas de las III Jornadas en Tecnología del Habla*, Valencia, del 17 al 19 de noviembre de 2004, Valencia, Departamento de Sistemas Informáticos y Computación, Facultad de Informática, Universidad Politécnica de Valencia.

SERENARI, M., L. DYBKJAER, U. HEID, M. KIPP y N. REITHINGER (2002): *Survey of existing gesture, facial expression and cross-modality coding schemes*, NITE, Natural Interactivity Tools Engineering, Deliverable D2.1., September 2002.
http://nite.nis.sdu.dk/deliverables/NITE-D2.1-sept02-F.pdf

SILVERMAN, K., M. BECKMAN, J. PITRELLI, M. OSTENDORF, C. WIGHTMAN, P. PRICE, J. PIERREHUMBERT y J. HIRSCHBERG (1992): «TOBI: A standard for labelling English prosody», en: *ICSLP 1992*, págs. 867-870.
http://ling.osu.edu/~tobi/ame_tobi/Silverman_etal1992.pdf

SJÖLANDER, K. y J. BESKOW (2000): «WaveSurfer - An open source speech tool», en: *ICSLP 2000 - INTERSPEECH 2000, Proceedings of the 6th International Conference on Spoken Language Processing*, 16 - 20 October, 2000, Beijin, China, págs. 464-467.
http://www.speech.kth.se/wavesurfer/wsurf_icslp00.pdf

SOSA, J.M. (2003): «La notación tonal del español en el modelo Sp-ToBI», en: P. Prieto (ed.), págs. 185-208.

URAGA, E. y C. GAMBOA (2004): «VOXMEX Speech Database : Design of a Phonetically Balanced Corpus», en: *LREC 2004*, págs. 1471-1474.

URAGA, E. y L. PINEDA (2002): «Automatic Generation of Pronunciation Lexicons for Spanish», en: A. Gelbukh (ed.) *CICLing 2002, Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, February 2003, Mexico City, Mexico, February 2002, Berlin: Springer, págs. 330-338.

VELDEN, J.G. VAN, D. LANGMANN y M. PAWLEWSKI (1996): *Specification of speech data collection over mobile telephone networks*, Version 2.3., SpeechDat LE2-401 Deliverable SD1.1.2/1.2.2., 14 October, 1996.
http://www.speechdat.org/speechdat/deliverables/public/SD112V23.DOC
http://www.sala2.org/sd112v23.rtf

VILLASEÑOR, L., A. MASSÉ y L.A. PINEDA (2000): «A Multimodal Dialogue Contribution Coding Scheme», en: *ISLE/EAGLES Workshop «Meta-Descriptions and Annotation Schemes for Multimodal/Multimedia Language Resources and Data Architectures and Software Support for Large Corpora»*, LREC 2000 Workshop, 29-30 May 2000, Athens, Greece.
http://www.mpi.nl/world/ISLE/documents/papers/villasenor_paper.pdf

VILLASEÑOR, L., M. MONTES, D. VAUFREYDAZ y J.F. SERIGNAT (2004): «Experiments on the construction of a phonetically balanced corpus from the web», en: A. Gelbukh (ed.) *CICLing 2004, Proceedings of the 5th International Conference on Intelligent Text Processing and Computational Linguistics*, 15-21 February, 2004. Seoul, Korea, Berlin, Springer, págs. 416-419.
http://ccc.inaoep.mx/~mmontesg/publicaciones/2004/PhoneticallyBalancedCorpus-cicling04.pdf

WEGENER, R., J.C. MARTIN, L. DYBKJAER, M.J. MACHUCA, N.O. BERNSEN, J. CARLETTA, U. HEID, S. KITA, J. LLISTERRI, C. PELACHAUD, I. POGGI, N. REITHINGER, G. VAN ELSWIJKS y P. WITTENBURG (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/D9.1-7.3.2002-F.pdf

WELLS, J.C. (1995a): *Computer-coding the IPA: a proposed extension of SAMPA*, Department of Phonetics and Linguistics, University College London.

http://www.phon.ucl.ac.uk/home/sampa/ipasam-x.pdf

WELLS, J.C. (1995b): *SAMPROSA: SAM Prosody Transcription*, Department of Phonetics and Linguistics, University College London. http://www.phon.ucl.ac.uk/home/sampa/samprosa.htm

WELLS, J.C. (1995c): *SAMPA for Spanish*, Department of Phonetics and Linguistics, University College London.

http://www.phon.ucl.ac.uk/home/sampa/spanish.htm

WELLS, J.C. (1997): «SAMPA computer readable phonetic alphabet», en: D. Gibbon, R. Moore y R. Winski (eds.), part IV, section B.

WELLS, J.C. (2003): «Phonetic symbols in word processing and on the web», en: *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 3-9 August, 2003, Barcelona, The ICPhS Organizing Committee, págs. 3105-3108.

http://www.phon.ucl.ac.uk/home/wells/ICPhS_18.pdf

WINSKI, R. (1997): *Definition of corpus, scripts and standards for fixed networks*, Version 4.1., SpeechDat LE2-401 Deliverable SD1.1.1., 22 January 1997.

http://www.speechdat.org/speechdat/deliverables/public/SD111V41.DOC

http://www.sala2.org/sd111v41.rtf

Proyectos y recursos citados en el texto

22 Language Corpus v1.2. http://cslu.cse.ogi.edu/corpora/22lang/

Anvil. http://www.dfki.de/~kipp/anvil/

BASURDE: Desarrollo de un sistema de diálogo oral en dominios restringidos. http://gps-tsc.upc.es/veu/basurde/Home.htm

DIHANA: Sistema de diálogo para el acceso a la información mediante habla espontánea en diferentes entornos. http://www.dihana.upv.es

EAGLES: Expert Advisory Group on Language Engineering Standards. http://www.ilc.cnr.it/EAGLES96/home.html

ELAN: EUDICO Linguistic Annotator. http://www.mpi.nl/tools/elan.html

ELDA: Evaluations and Language resources Distribution Agency. http://www.elda.org/

EMU Speech Database System: http://emu.sourceforge.net/

EUROM 1 Spanish Database. http://gps-tsc.upc.es/veu/LR/LR_EuromI.html

FAME: Facilitating Agents in Multicultural Exchange. http://isl.ira.uka.de/fame/

INTERFACE: Multimodal Analysis/Synthesis System for Human Interaction to Virtual and Augmented Environments. http://ligwww.epfl.ch/~thalmann/interface.html

International Phonetic Alphabet. http://www.arts.gla.ac.uk/IPA/ipachart.html

ISLE: International Standards for Language Engineering - Natural Interaction and MultiModality Working Group. http://isle.nis.sdu.dk/

LC-STAR: Lexica and Corpora for Speech-to-Speech Translation Components. http://www.lc-star.com/

LDC: Linguistic Data Consortium. http://www.ldc.upenn.edu/

MATE: Multilevel Annotation, Tools Engineering. http://mate.nis.sdu.dk/

MES: Motif Environement for Speech. http://aune.lpl.univ-aix.fr/ext/projects/mes_signaix.htm/

MOMEL: Modélisation de mélodie. http://www.lpl.univ-aix.fr/~hirst/software.html
MULTEXT: Multilingual Text Tools and Corpora. http://www.lpl.univ-aix.fr/projects/multext/
Multilanguage Telephone Speech Corpus v1.2. http://cslu.cse.ogi.edu/corpora/mlts/
NIST: National Institute of Standards and Technology, Speech Group. http://www.nist.gov/speech/index.htm
NITE: Natural Interactivity Tools Engineering. http://nite.nis.sdu.dk/
Praat. http://www.praat.org
SALA: SpeechDat across Latin America. http://www.sala2.org/
SAMPA Computer Readable Phonetic Alphabet. http://www.phon.ucl.ac.uk/home/sampa/home.htm
SAMPA European and Latin American Spanish Allophone Set. http://www.sala2.org/SALASAMPA.rtf
SAMPA for Spanish. http://www.phon.ucl.ac.uk/home/sampa/spanish.htm
SAMPROSA: SAM Prosody Transcription. http://www.phon.ucl.ac.uk/home/sampa/samprosa.htm
SIL Doulos IPA Font, SIL International. http://scripts.sil.org/DoulosSILfont
SpeechDat Car Spanish. http://gps-tsc.upc.es/veu/LR/LR_SPDCAR.html
SpeechDat Spanish FDB 4000 speakers. http://gps-tsc.upc.es/veu/LR/LR_SPD_FDB.html
SpeechDat. http://www.speechdat.org/
SpeechWorks, Scicon R&D, Inc.. http://www.sciconrd.com/pworks.htm
Sp-ToBI: Spanish Tones and Break Indices. http://www.ling.ohio-state.edu/~tobi/sptobi/spanish.html
Tlatoa Common Questions Corpus. http://www.udlap.mx/~sistemas/tlatoa/corpora/telephone_collection_sp.html
Tlatoa/OGI Spanish TTS Corpus. http://info.udlap.mx/~sistemas/tlatoa/corpora/tts_corpus_sp.html
ToBI: Tone and Break Indices. http://www.ling.ohio-state.edu/~tobi/
Transana. http://www.transana.org/
Transcriber. http://www.etca.fr/CTA/gip/Projets/Transcriber/
VAESS: Voices, Attitudes and Emotions in Speech Synthesis. http://www.speech.kth.se/speech/proj/vaess.html
Verbmobil. http://verbmobil.dfki.de/
VILE: Estudio acústico de la variación inter e intra locutor en español. http://liceu.uab.es/~joaquim/VILE.html
WaveSurfer. http://www.speech.kth.se/wavesurfer/index.html
Worldbet. ASCII phonetic symbols for the world's languages. http://www.ling.gu.se/~jimh/courses/ipa.ps http://www.ling.gu.se/~jimh/courses/ipa.recog.unicode.ps
X-SAMPA. Computer coding of the IPA: A proposed extension of SAMPA. http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm