Diseño de corpus textuales y orales

Joan Torruella y Joaquim Llisterri

Seminari de Filologia i Informàtica
Departament de Filologia Espanyola
Universitat Aut noma de Barcelona

A la memoria de nuestro buen amigo Giovanni Pontiero.

1. Introducción

Cada vez parece más evidente la conveniencia de utilizar recursos informáticos en las investigaciones humanísticas. Pero para poder utilizar estos recursos es necesario disponer de un material donde aplicarlos; este material, en el caso de la filología, son los textos, orales o escritos, y los documentos que los contienen, los cuales, debidamente recopilados, forman los llamados corpus.

Actualmente, en muchas ramas de las humanidades, y sobretodo en lingüística aplicada, se pretende trabajar con datos reales y lo más exhaustivos posibles que permitan reproducir con la máxima fidelidad las características del objeto de estudio. Esto implica que, de algún modo, hay que recopilar, en cantidades más o menos grandes, muestras de los elementos que constituyen la realidad que se quiere observar. El auge que últimamente ha tenido la aplicación de la informática y su inevitable presencia en cualquier campo de la investigación ha facilitado enormemente las tareas mecánicas de recopilación y organización en formato electrónico de los textos, lo cual ha provocado que el investigador se pueda encontrar delante de cantidades considerables de documentos que aportan un número de datos tan grande que sólo una codificación, ordenación y organización de estos datos en la proporción adecuada pueden salvarlo del naufragio en un mar inmenso de información. De ahí que, en este capítulo, más que describir investigaciones concretas que se han llevado a cabo en el área de los corpus o a partir de ellos, presentamos las pautas para obtener un corpus suficientemente organizado y representativo de la realidad que quiera reflejar, para que pueda ser explotado con ciertas garantías de éxito[1].

Ya J. Svartvik (1992) señaló que la lingüística basada en los corpus hacía posible nuevas aproximaciones a viejos problemas, y no solamente esto sino que, en muchos casos permite poner en el terreno de las afirmaciones ideas que antes solo eran conjeturas o especulaciones provenientes de impresiones más o menos fundadas de los lingüistas. Una característica importante de los corpus es que estan compuestos por datos reales y, por lo

[1] Para una presentación general de la lingüística de corpus véase, por ejemplo, Leech (1991), Leech y Fligelstone (1992) o McEnery y Wilson (1996).

tanto, sus resultados son empíricos, a diferencia de otras metodologías de análisis lingüístico en las que se parte de hipótesis más intuitivas. La función principal de un corpus, tanto textual como oral, es establecer la relación entre la teoría y los datos; el corpus tiene que mostrar a pequeña escala cómo funciona una lengua natural; pero para ello es necesario que esté diseñado correctamente sobre unas bases estadísticas apropiadas que aseguren que el resultado sea efectivamente un modelo de la realidad. Si el corpus tiene que ser un modelo de la realidad lingüística, o de una parte de esta realidad, es necesario que sea neutro, o sea, que recoja muestras proporcionales de todos sus aspectos (niveles, temáticas, registros, etc.). En la medida en que un corpus sea neutro, es decir, no marcado, se podrá explotar posteriormente para trabajos y enfoques diferentes: fonéticos, fonológicos, morfológicos, sintácticos, semánticos, pragmáticos, etc., siendo constantemente un producto actualizable y reutilizable, dos conceptos importantísimos de la investigación de hoy, ya que si la tarea de confección de un corpus es considerable, a pesar de la ayuda informática, lo mínimo que hay que asegurar es que el resultado sea rentable, y lo será en la medida en que pueda ser utilizado en diversos medios y para diversos fines. De todos modos, hay que aceptar el hecho de que la neutralidad es una tendencia y no una realidad ya que siempre dirigimos la mirada o el pensamiento hacia aquello que, consciente o inconscientemente, queremos ver o demostrar; "no deberíamos olvidar que lo que observamos no es la naturaleza misma, sino la naturaleza determinada por la índole de nuestras preguntas"[2].

Por eso hay que tener siempre presente que un corpus nunca puede ser la realidad sino solamente un modelo de ésta, modelo que debería mostrar sus aspectos más destacados y más característicos. Cuanto más grande sea el corpus y el número de niveles, tipologías, etc. de textos que lo integren más posibilidades habrá de asegurar la presencia de todos los aspectos de la lengua y, por lo tanto, de acercarse a la realidad. Pero un corpus siempre tiene que ser selectivo ya que no es posible (y de serlo tampoco sería rentable), recopilar todo lo escrito y/o hablado de una lengua, y, de hecho, operativamente, es preferible un corpus bien seleccionado y representativo a un corpus exhaustivo, que lo quiera recoger todo. El carácter selectivo de los corpus puede limitar algunas veces las posibilidades de extraer conclusiones, ya que, por ejemplo, en la lista de frecuencias de las unidades léxicas presentes en cualquier corpus, por grande que este sea, un número bastante elevado de unidades (la mitad aproximadamente) tienen frecuencia absoluta de aparición 1, con lo cual no es posible extraer según que tipo de informaciones referentes a estas unidades ni poder explicar su funcionamiento dentro de la lengua. Y lo mismo podríamos decir de las palabras con un índice de frecuencia absoluta de aparición superior a 1 pero sin llegar a ser lo suficientemente grandes como para permitir deducir generalizaciones.

Para que los corpus faciliten la extracción de datos homogéneos y cuantificables de manera que permitan elaborar teorías empíricas, es necesario restringir las diferentes ocurrencias léxicas a ocurrencias formales comunes (unidades estandarizadas); para ello es

[2] Cita tomada de Marina (1993: 38).

necesario reducir las variantes a invariantes[3]. Y no debemos entender estas variantes solo como las puramente gráficas, las de carácter fonético o las de naturaleza diatópica, sino también las producidas por la polisemia de las lenguas (en muchas de ellas la mitad de las palabras tienen más de una acepción). Otro paso también necesario de cara a reducir a un común denominador las diferentes formas flexivas que adquieren las palabras cuando son tratadas únicamente como cadenas de caracteres delimitadas entre espacios en blanco, es agrupar bajo de un lema todas sus formas flexionadas. Pero todos estos procesos suponen ya una teoría previa de la morfología.

Todo esto ha hecho que la creación y el mantenimiento o actualización de los corpus se haya convertido en una ciencia interdisciplinar en la que no solamente tienen que intervenir los lingüistas sino también historiadores, sociolingüistas, matemáticos, informáticos, teóricos de la literatura, etc. Decidir el tamaño que tiene que tener un corpus textual u oral y cada una de las muestras que van a configurarlo para que éste sea un reflejo de la lengua que pretende representar no es nada fácil, como tampoco lo es definir las diferentes etapas diacrónicas posibles, la variedad temática que ha de contemplar o la conveniencia de trabajar con documentos enteros o con fragmentos de cada uno de ellos. Establecer los documentos y las ediciones más representativas para incluirlas en el corpus puede ser algo muy subjetivo si no se hace siguiendo algún criterio mínimamente imparcial: ¿se han de escoger los textos más leídos?, ¿los más reconocidos?,
¿seleccionados al azar?, etc.; ¿quién se atreve a priorizar obras de tanto prestigio como el *Quijote* o *Cien años de soledad* frente a otras con tanta difusión como pueden ser las de Corín Tellado, J.J. Benítez o Vizcaíno Casas?, y ¿con qué criterio?

Actualmente existe un gran número de corpus, muy variados por lo que respecta a la extensión, al diseño y a las finalidades[4]. El hecho es que los corpus informatizados han demostrado ser unas herramientas excelentes para muchos tipos de investigaciones; principalmente en el campo de la investigación lingüística porque, como ya se ha dicho, proporcionan bases mucho más reales para el estudio de las lenguas que los métodos intuitivos tradicionales. A partir de los corpus podemos disponer de bases muy provechosas para comparar diferentes variedades de una lengua o para explotar sus aspectos cuantitativos y probabilísticos. Efectivamente, los corpus informatizados han venido a dar un nuevo impulso a los estudios descriptivos de los diferentes aspectos de la lengua: prosodia, léxico, morfología, sintaxis, historia de la lengua, etc[5].

A parte de estas cuestiones más generales, los corpus informatizados han influido y cambiado bastante los métodos de investigación e, incluso, han propiciado el nacimiento

[3] Para este tema en concreto y otros relacionados con los corpus, véase el excelente capítulo de Blecua (en prensa).

[4] Algunos inventarios de corpus existentes pueden encontrarse en Cole (ed.) (1996), Edwards (1993), McEnery y Wilson (1996) y en los catálogos de ELRA (*European Language Resources Association*) <http://www.icp.grenet.fr/ELRA/catalog.html> o de LDC (*Linguistic Data Consortium*) < http://www.ldc.upenn.edu/ldc/catalog/index.html>. Véanse también Taylor *et al.* (1991) para el inglés, Fernández y Llisterri (1996a) o Llisterri (1996) para el español y Badia *et al.* (1994) para el catalán.

[5]

Una muestra reciente de ello la constituyen los trabajos recogidos en Thomas y Short (eds.) (1996).

de nuevas tendencias lingüísticas. Muchos trabajos que antes tenían que hacerse a mano, empleando mucho tiempo y esfuerzo leyendo y repasando textos para encontrar datos concretos que sirvieran para demostrar nuestras hipótesis, hoy, con la ayuda de la informática, se pueden hacer no solamente con menos tiempo sino también más ordenada y exhaustivamente, o sea, con mayor eficacia y eficiencia. Los avances más significativos en el campo de la lingüística de corpus se han producido en el área de la creación de modelos probabilísticos de la lengua y como pruebas para verificar estos modelos; ello ha permitido avanzar significativamente en el campo de los análisis gramaticales automáticos de textos, tanto en sus aspectos morfológicos (*tagging*) como en sus aspectos sintácticos (*parsing*). Pero últimamente se han producido avances considerables en áreas más aplicadas, como la de la traducción automática o la del reconocimiento y síntesis del habla.

La ventajas de trabajar con corpus informatizados, sobre todo con aquellos que están anotados, es tan grande, que está obligando a los lingüistas "tradicionales" a trabajar conjuntamente con lingüistas computacionales. La finalidad última, sin embargo, es siempre la misma: entender mejor cómo funciona el lenguaje humano, a pesar de que la finalidad inmediata pueda ser obtener datos para preparar un curso de lengua para extranjeros, para confeccionar un programa de traducción automática, para construir un conversor de texto a habla, etc[6].

La lexicografía y la terminología son dos de los campos de investigación y de estudio que más se benefician de las informaciones que los corpus textuales y los corpus de lengua oral aportan. Éstos son de gran ayuda para configurar el lemario de los diccionarios (tanto para incluir nuevas palabras como para excluir las desusadas), así cómo para separar las distintas acepciones de cada lema, para detectar las palabras co-ocurrentes, las combinaciones sintácticas, etc. Los corpus también proporcionan material muy útil para trabajar sobre fraseología, la detección de neologismos y la obtención de ejemplos reales susceptibles de aparecer en los diccionarios[7].

Este método de trabajo también resulta muy productivo en el campo de la estadística lingüística donde se utiliza para establecer índices de frecuencias tanto de palabras, morfemas, sílabas, letras, etc., como de combinaciones léxicas de distinta naturaleza. Así se pueden definir las reglas combinatorias de los formantes léxicos, el grado de vitalidad de los elementos de formación de palabras, la frecuencia de aparición de diferentes tipos de vocablos (tecnicismos, barbarismos, neologismos, etc.) o de diferentes niveles del lenguaje (vulgar, culto, literario, etc.), datos, estos últimos, muy interesantes no solo para los estudios lexicógrafos sino también para los estudios sociolingüísticos y estilísticos.

[6] Un útil resumen de las aplicaciones de los corpus se encuentra en el capítulo 4 de McEnery y Wilson (1996). Pueden encontrarse ejemplos específicos en Aarts y Meijs (eds.) (1986), (1990), Oostdijk y de Haan (eds.) (1994) y Svartvik (ed.) (1992). Específicamente dedicadas al inglés son las recopilaciones de Aarts y Meijs (eds.) (1984), Aarts *et al*. (eds.) (1993), Aijmer y Altenberg (eds.) (1991), de Haan y Oostdijk (eds.) (1993), Fries *et al*. (eds.) (1994), Johansson (ed.) (1982), Johansson y Stenström (eds.) (1991), Kytö *et al.* (eds.) (1988), Leitner (ed.) (1992), Meijs (ed.) (1987). Para el español, véase Alvar y Villena (Coord.) (1994) y Sánchez *et al*. (1995).

[7]

Uno de los ejemplos más clásicos de la aplicación de los corpus a la lexicografía lo constituye el *Collins-COBUILD English Language Dictionary* (Sinclair (ed.), 1987). Puede encontrarse más información sobre este proyecto en <http://titania.cobuild.collins.co.uk/>.

En el terreno de la gramática histórica y la historia de la lengua, los corpus proporcionan datos referentes a la formación de palabras, a los cambios de significado producidos en un vocablo, a las diferentes áreas de utilización de una voz, a las evoluciones formales de una palabra, a la introducción de palabras no normativas en la lengua, etc[8].

Otro campo en el que los corpus aportan grandes ventajas es el de la confección de herramientas lingüísticas informatizadas. Una de las más importantes es la de los diccionarios-máquina, de usos tan diversos como la corrección de textos informatizados o la segmentación de las palabras por sílabas. Estas herramientas son importantísimas para la traducción automática y otras tareas basadas en el tratamiento automático del lenguaje[9].

En el campo de la fonética, los corpus constituidos por grabaciones de laboratorio son herramientas imprescindibles para el estudio experimental del habla, mientras que los que contienen registros menos formales son necesarios para la caracterización de diversos estilos. En el ámbito de las tecnologías del habla, las bases de datos orales proporcionan datos importantes para la modelización de los fenómenos segmentales y suprasegmentales en la conversión de texto a habla y son esenciales para el entrenamiento y la validación de los sistemas de reconocimiento y de diálogo en entornos de comunicación persona máquina, cuyas aplicaciones se extienden desde la oferta de servicios telefónicos automatizados hasta las ayudas para personas con discapacidades.

Los corpus también pueden proporcionar elementos muy útiles en el campo de la enseñanza de lenguas[10], sobre todo a la hora de preparar materiales o ejercicios de trabajo en clase basados en un uso real de la lengua. Del contenido de los corpus puede desprenderse información tanto de uso (palabras y construcciones más frecuentes en los libros de texto y lecturas recomendadas en relación con los materiales auténticos) como de corrección de barbarismos o malos usos lingüísticos (errores más repetidos, construcciones no normativas, léxico mal usado, grafías incorrectas, etc.). La recopilación de corpus de producciones de estudiantes de lengua extranjera constituye también una fuente de datos sobre la interferencia entre la primera y la segunda lengua en todos los niveles del análisis lingüístico y una base empírica importante para el análisis de errores y de las estrategias comunicativas de los alumnos.

En cuanto a las utilidades de los corpus en otros campos de las humanidades que no sean los estrictamente lingüísticos cabe mencionar las posibilidades que ofrecen para los estudios históricos, para los de la teoría de la literatura, etc. Si los textos que componen un corpus están asociados a una documentación detallada de sus rasgos externos: fecha, tema, región, edad del autor, *estatus* social, sexo, etc., éstos pueden convertirse en fuente de datos para aquellas personas interesadas en los aspectos de contenido textual los

[8] Sobre las aplicaciones de los corpus a la diacronía véanse, por ejemplo, los estudios reunidos en Kytö *et al.* (eds.) (1994) o en Rissanen *et al*. (eds.) (1993).

[9]

Para un tratamiento más detallado de los usos de los corpus en la lingüística computacional véase el capítulo 5 de McEnery y Wilson (1996); trabajos más específicos pueden encontrarse en Souter y Atwell (eds.) (1993).

[10] Véase, por ejemplo, Knowles (1990) o Mindt (1996).

historiadores, por ejemplo, pueden seguir la evolución de opiniones e ideas mediante el estudio de palabras o frases asociadas a ellas.

En la sociolingüística, aunque usando parámetros diferentes de los utilizados por los historiadores, también se pueden obtener de los corpus datos de gran utilidad; al contrario que a los estudiosos de la historia, a los sociolingüistas no les interesa tanto el tema del texto o el nombre del autor como la clase social, el sexo o el nivel cultural del receptor. Estrechamente relacionada con el uso de corpus en la sociolingüística está la utilización de los mismos como base de estudios dedicados a la diferenciación entre registros o estilos por ejemplo entre la lengua escrita y la oral o entre diversos 'géneros' como la correspondencia privada, el discurso jurídico, político, publicitario o religioso, incluyendo incluso trabajos sobre las características de los mensajes de correo electrónico - asociados a variaciones en la situación de comunicación y a dimensiones como el grado de formalidad, el carácter público o privado, etc [11]. Estos trabajos entroncan directamente con los realizados desde la perspectiva del análisis del discurso, encaminados a establecer tipologías textuales.

La psicolingüística puede también verse beneficiada por el uso de corpus, especialmente en campos como el análisis de los errores de producción del habla o el desarrollo del lenguaje infantil[12]. El análisis de las patologías del lenguaje y del habla requiere igualmente colecciones sistemáticas de muestras recogidas de personas que presentan transtornos de la comunicación.

También los estudiosos de la literatura pueden tener en los corpus una buena herramienta para sus investigaciones. En el campo de la estilística, por ejemplo, los corpus pueden ayudar a definir los trazos que caracterizan distintos estilos literarios o, en el terreno de la estilometría, los análisis estadísticos del uso de las palabras en los textos pueden dar luz a problemas de adscripción de trabajos de dudosa autoría.

2. ¿QUÉ ES UN CORPUS LINGÜÍSTICO (INFORMATIZADO)?

Durante los últimos años ha habido, tanto en América como en Europa y Japón, un gran crecimiento del interés en la creación y explotación de corpus lingüísticos como parte de la infraestructura para el desarrollo de aplicaciones encaminadas al procesamiento del lenguaje. El tratamiento estadístico de los datos que facilitan los corpus ha demostrado ser eficaz para encontrar la solución a algunos problemas tradicionales de la lingüística computacional, de la traducción automática, etc. El auge que ha tomado esta disciplina ha hecho que actualmente en casi todos los centros de investigaciones lingüísticas se esté trabajando en la confección de algún tipo de corpus.

Pero, ¿qué es un corpus? ¿Entendemos todos lo mismo cuando hablamos de corpus?

[11] Una revisión de los trabajos sobre registro en esta línea puede encontrarse en Atkinson y Biber (1994). Se enmarcan también en esta perspectiva Biber y Finegan (1991) o Biber (1990).

[12] En este campo es especialmente relevante el proyecto CHILDES (MacWhinney, 1991) sobre el cual puede obtenerse más información en <http://poppy.psy.cmu.edu/childes/childes.html>.

Según J. Sinclair, uno de los grandes especialistas en el campo de los corpus modernos, un *corpus* es:

> A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language (Sinclair, 1994:4)[13].

Según esta definición la informática no tiene que ver con el concepto de "corpus", y, de hecho, así es. Pero hoy en día la informática facilita tanto la organización y la explotación de grandes cantidades de datos que sería impensable crear un corpus prescindiendo de este medio o herramienta. Por esto, hoy más que hablar de *corpus* hay que hablar de *corpus informatizados* ya que son dos conceptos íntimamente ligados.

Así, según el mismo J. Sinclair, un *corpus lingüístico informatizado* es:

> *... a corpus which is encoded in a standardised and homogenous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance* (Sinclair, 1996:4)[14].

2.1. COLECCIONES DE TEXTOS

En el campo de la lingüística, la palabra *corpus* es una palabra algo ambigua y que actualmente se utiliza en un sentido general para referirse a cualquier tipo de recopilación de textos. En realidad, para ser más exactos, en el ámbito de la recopilación de textos hay que distinguir, según el grado de especificación en los criterios de selección, al menos entre tres tipos diferentes de recopilaciones:

- Archivo/colección (informatizado) (*Archive/Collection*).- Es un repertorio de textos en soporte informático sin buscar ningún tipo de relación entre ellos.
- Biblioteca de Textos Electrónicos (*Electronic text library*).- Es una colección de textos en soporte informático, guardados en un formato estándar, siguiendo ciertas normas de contenido, pero sin un criterio riguroso de selección.
- Corpus Informatizado (*Computer corpus*).- Es una recopilación de textos seleccionados según criterios lingüísticos, codificados de modo estándar y homogéneo, con la finalidad de poder ser tratados mediante procesos informáticos y destinados a reflejar el comportamiento de una o más lenguas.

Los dos primeros tipos de recopilaciones no implican una selección o una ordenación hecha siguiendo criterios lingüísticos, mientras que los corpus sí. Estos criterios lingüísticos pueden ser a) externos o b) internos (Sinclair, 1996:5).

[13]
(Los subrayados son nuestros). [14]
(Los subrayados son nuestros).

a) son *externos* cuando hacen referencia a datos de los autores, a los medios de transmisión utilizados, al nivel social de los participantes, a la función comunicativa de los textos, etc.
b) son *internos*, cuando hacen referencia a patrones lingüísticos presentes en los textos. (Sinclair, 1996:4)

2.2. NIVELES EN LOS CORPUS

En una selección de textos destinada a constituir un corpus propiamente dicho podemos encontrar diferentes niveles: corpus, subcorpus y componentes.

- Corpus.- Un corpus es un conjunto homogéneo de muestras de lengua de cualquier tipo (orales, escritos, literarios, coloquiales, etc.) los cuales se toman como modelo de un estado o nivel de lengua predeterminado. El conjunto de enunciados incluidos en un corpus, una vez analizados, debe permitir mejorar el conocimiento de las estructuras lingüísticas de la lengua que representan.
- Subcorpus.- Suele ser una selección estática de textos, derivada de un corpus normalmente más general y complejo, el cual está dividido en grupos de muestras textuales más específicas; pero también puede ser una selección dinámica de textos de un corpus en crecimiento: un número determinado de textos destinados a aumentar algún apartado de un corpus general.
- Componente.- Es una colección de muestras de un corpus o de un subcorpus, las cuales responden a un criterio lingüístico específico muy concreto. Los componentes reflejan un tipo determinado de lengua. Sobre todo los corpus, pero también los subcorpus, son muy heterogéneos, mientras que los componentes son muy homogéneos.

2.3 CORPUS TEXTUALES Y CORPUS ORALES

Llegados a este punto, parece conveniente detenerse brevemente en la distinción entre los llamados ‘corpus textuales’ y los ‘corpus orales’. Mientras que en el caso de los primeros es claro que constituyen muestras de la lengua escrita, los segundos pueden consistir tanto en transcripciones ortográficas de la lengua hablada como en grabaciones acompañadas de la correspondiente transcripción. La procedencia de las grabaciones suele ser muy diversa: desde las que se realizan en laboratorios de fonética con materiales altamente controlados hasta las obtenidas en entrevistas espontáneas o las recogidas de los medios de

comunicación, incluyendo también las interacciones ficticias usadas en el diseño de los sistemas de diálogo persona-máquina.

Mientras que en el campo de la lingüística de corpus existe una tendencia a considerar como corpus orales (*spoken corpora*) las transcripciones ortográficas del habla, tanto en fonética como en tecnologías del habla difícilmente se concibe un corpus que no vaya acompañado del correspondiente registro sonoro en formato digital (*speech corpus*). Sin embargo, la necesidad de obtener modelos estadísticos de la lengua en el desarrollo de sistemas de reconocimiento pensados para aplicaciones como el dictado automático ha llevado a un uso cada vez más frecuente de los corpus textuales y de las transcripciones del registro oral espontáneo en este ámbito. Por otro lado, el interés por los aspectos prosódicos del discurso y la conversación hace que desde la lingüística de corpus tradicional surja la necesidad de disponer de grabaciones sincronizadas temporalmente con la transcripción, sea ésta ortográfica, fonética o fonológica.

En el presente capítulo, utilizaremos 'corpus oral' para referirnos a todo tipo de materiales, tanto transcripciones como grabaciones, en los que se recoge la lengua hablada[15]. Nos referiremos también a 'textos' como elementos integrantes de un corpus, tanto si constituyen material originariamente escrito como si provienen de transcripciones de la lengua oral.

3. Clasificación de los corpus

3.1. Criterios generales para la clasificación de los corpus

Los diferentes tipos de corpus se pueden clasificar de diferentes maneras en función de los parámetros que se quieran utilizar: 3.1.1. según el porcentaje y la distribución de los diferentes tipos de textos que lo componen; 3.1.2. según la especificidad de los textos que lo componen; 3.1.3 según la cantidad de texto que se recoge en cada documento; 3.1.4. según la codificación y las anotaciones añadidas a los textos; 3.1.5. según la documentación que le acompañe.

En principio, un corpus bien estructurado ha de responder, aunque sea por defecto, a algún parámetro de cada uno de estos grupos. Veamos ahora con más detalle en qué consisten los criterios mencionados.

3.1.1. Según el porcentaje y la distribución de los diferentes tipos de texto

Los corpus pueden clasificarse según la distribución y el porcentaje escogido de los diferentes tipos de texto que lo componen. Según estos parámetros tenemos:

1. *Corpus grande*.- Corpus que no se plantea el límite del volumen de textos que ha de recoger o que, si se lo plantea, lo cuantifica en un número de palabras muy elevado sin tener en cuenta cuestiones de equilibrio, de representatividad, etc.

Esta característica es, en muchos casos, ambigua, ya que se habla de *corpus grandes* pero sin precisar las dimensiones en número de unidades léxicas que un corpus ha de tener para ser considerado como tal. El valor por defecto de los diferentes tipos de

[15] Una elaboración más detallada de la distinción entre *speech corpora* y *spoken corpora* puede encontrarse en Llisterri (1996b).

corpus en cuanto a su extensión es "grande" por oposición a corpus cuantitativamente más pequeños como pueden ser los *corpus monitor*, los *corpus piramidales*, etc., los cuales, a pesar de que también pueden ser muy extensos, tienen que tener controlado el volumen de cada tipo de textos que los componen. De todos modos, el volumen de los corpus crece constantemente, sobre todo gracias a las facilidades informáticas para su recopilación, manipulación y explotación, por lo que el término "corpus grande" se ha de entender más en el sentido de opuesto a otros tipos de corpus voluntariamente delimitados en su extensión que en un sentido de cantidad.

2. *Corpus equilibrado*.- Corpus que contiene diferentes variedades de textos distribuidos cuantitativamente en proporciones parecidas para cada variedad.
3. *Corpus piramidal*.- Corpus en que sus componentes, o sea sus textos, están distribuidos en diversos estratos o niveles: un primer estrato que recoge pocas variedades temáticas pero con muchos textos en cada variedad; un segundo estrato que recoge mayor variedad de textos pero menos cantidad en cada una de ellas; un tercer estrato compuesto por muchas variedades pero con pocos textos en cada variedad; y así hasta un número de estratos opcional.
4. *Corpus monitor*.- Este tipo de corpus es consecuencia de la gran cantidad de palabras que últimamente están incluyendo los corpus. Las grandes dimensiones de los corpus hacen que sean difíciles de controlar y de explotar. Para evitarlo, los corpus monitor quieren tener un volumen textual constante pero en continua actualización. El conjunto de textos que lo componen se va renovando cada cierto tiempo de manera que siempre se van incluyendo nuevos textos al mismo tiempo que se van excluyendo otros, consiguiendo de este modo un corpus vivo y dinámico como lo es la propia lengua. Normalmente la inclusión y exclusión de textos se hace siguiendo pautas temporales (se incluyen textos del último año y se excluyen los del primero) y conservando debidamente ordenados los textos que se van excluyendo, de manera que podemos llegar a tener un buen material para construir un *corpus diacrónico*, ya que podremos disponer de diversos grupos de textos con más o menos las mismas proporciones y las mismas características pero representantes de momentos sucesivos de la lengua. De este modo se pueden establecer las frecuencias de distribución de las palabras en diversas etapas cronológicas, identificar neologismos, palabras que entran en desuso, nuevas acepciones de palabras ya existentes, etc.

   A lo largo del tiempo, la distribución de los distintos grupos y componentes de un corpus monitor va cambiando porque siempre van apareciendo nuevos temas y nuevas fuentes,

por lo que las distintas proporciones se han de ir ajustando para poder reflejar mejor la realidad lingüística de cada momento.

5 *Corpus paralelo*.- Es una colección de textos traducidos a una o varias lenguas. El más sencillo es el que consta del original y su traducción a otra lengua. La dirección de la traducción no es necesario que sea constante, un corpus paralelo puede contener tanto textos traducidos de la lengua A a la lengua B como textos traducidos de la lengua B a la lengua A. Este tipo de corpus es de gran utilidad sobre todo en el campo de la traducción, y principalmente de la traducción automática, ya que los programas suelen trabajar con datos probabilísticos que sólo pueden obtenerse a partir de los corpus.

6. *Corpus comparables*.- Son corpus que seleccionan textos parecidos en cuanto a sus características en más de una lengua o en más de una variedad. Una de las principales finalidades de este tipo de corpus es poder comparar el comportamiento de diferentes lenguas o de diferentes variedades de una lengua en circunstancias de comunicación parecidas pero evitando las inevitables distorsiones lingüísticas introducidas en las traducciones recogidas en los corpus paralelos.
7. *Corpus multilingües*.- J. Sinclair sugiere que cuando se recopilan textos de diferentes lenguas sin que sean traducciones unos de otros y sin compartir criterios de selección, como lo hacen los textos que componen un corpus comparable, habría que hablarse de corpus multilingües.
8. *Corpus oportunista.*- Corpus que recoge textos que encuentra disponibles sin seguir ningún criterio de selección. Esto normalmente está motivado por la poca disponibilidad de textos en soporte electrónico (aun que cada vez se pueden encontrar en mayor cantidad) y por el elevado número de palabras necesarias para poder realizar muchos trabajos de investigación y la falta de recursos para obtenerlas. En realidad, de acuerdo con lo dicho en el apartado anterior, en este caso no se debería hablar de *Corpus Oportonista* sino que se debería hablar de *Archivo de Textos Informatizado* o de *Biblioteca de Textos Electrónicos*.

3.1.2. Según la especificidad de los textos

Otra clasificación que se puede hacer de los corpus es en función de la especificidad de los textos que lo componen. Atendiendo a este parámetro podemos definir cuatro tipos:

1. *Corpus general.*- Corpus que, al pretender reflejar la lengua común en su ámbito más amplio, se interesa por recoger cuantos más tipos de géneros mejor. Este tipo de corpus es útil para describir la lengua común de una colectividad, el lenguaje que utilizan los hablantese en situaciones comunicativas normales.
2. *Corpus especializado*.- Se opone al corpus general. El corpus especializado recoge textos que puedan aportar datos para la descripción de un tipo particular de lengua. El corpus especializado es diferente al corpus que contempla una o más variedades de la lengua general (*subcorpus*); un corpus que recoja conversaciones de la calle no es un corpus

especializado, como tampoco lo es uno que recoja el lenguaje de los periódicos; sí que lo sería, por ejemplo, un corpus que solo recogiera textos poéticos.

3. *Corpus genérico*.- Corpus condicionado por el género de los textos que contiene, interesándose solo por algunos de ellos; por ejemplo, una recopilación de textos de revistas científicas especializadas o la selección de textos poéticos.
4. *Corpus canónico*.- Corpus formado por todos los textos que configuran lo obra completa de un autor, independientemente de los géneros.
5. *Corpus periódico* o *cronológico*.- Corpus que recoge textos de unos años determinados o de unas épocas concretas.
6. *Corpus diacrónico*.- Corpus que incluye textos de diferentes etapas temporales sucesivas en el tiempo con el fin de poder observar evoluciones en la lengua.

3.1.3. Según la cantidad de texto que se recoge de cada documento

También se pueden clasificar los corpus según la cantidad de texto que se escoja de cada documento para cada muestra. Atendiendo a este criterio los corpus se pueden dividir en:

1. *Corpus textual (Whole text corpus)*.- Corpus que recoge íntegramente todos los textos de los documentos que lo constituyen. Se entiende como textos enteros las series de frases y/o párrafos coherentes, homogéneos estilísticamente y completos en sí mismos. Las novelas, por ejemplo, son un prototipo de texto que cumple estos requisitos, pero hay otros tipos de documentos que también se adaptan a esta definición. Atkins y otros consideran como un texto entero las recopilaciones de pequeños anuncios de periódico o colecciones de poemas cortos de un mismo autor. A veces incluso todos los artículos de un periódico o de una revista se han considerado como un solo texto, aunque es más razonable considerar como un solo texto los diversos artículos de una misma sección (economía, deportes, editoriales, etc.) aparecidos en diversos números de la misma publicación. El caso de los textos que aparecen en la sección de "cartas al director", textos que por su procedencia pueden ser muy interesantes, es un caso algo especial que los editores del corpus deberán considerar.
2. *Corpus de referencia (Reference corpora)*.- Corpus formado por fragmentos de los textos de los documentos que lo constituyen. En este caso no interesa tanto el texto en sí sino el nivel de lengua que representan. En este tipo de corpus son muy importantes los aspectos de equilibrio y representatividad en la selección de los fragmentos.
3. *Corpus léxico (Samples corpus)*.- Corpus que recoge fragmentos de textos muy pequeños y de longitud constante de cada documento. En este caso el interés de los diseñadores del corpus está en el léxico.

3.1.4. Según la codificación y la anotación

También se pueden clasificar los corpus atendiendo a las etiquetas descriptivas y analíticas que se han usado en la codificación de los textos. Según estos criterios los corpus serán:

1. *Corpus simple (o no codificado ni anotado)*.- Corpus que ha sido guardado en formato neutro (ASCII, también llamado *plain text*), y sin codificación para ninguno de sus aspectos.
2. *Corpus codificado o anotado*- Corpus formado por textos a los cuales se les ha añadido, ya sea manual o automáticamente, etiquetas declarativas de algunos elementos estructurales de los documentos (indicación de título, de principio de capítulo, de cambio de lengua, etc.) - codificación- o etiquetas analíticas de algunos aspectos lingüísticos (indicación de frase subordinada, de aspectos pragmáticos, etc.). - anotación
   - [16] De todos modos es importante que las etiquetas usadas para codificar y anotar los textos sean siempre extratextuales, de manera que se puedan reconocer y, si es necesario, eliminar fácilmente. También es importante que se usen sistemas de codificación estándares para asegurar la transportabilidad y reusabilidad de los textos.[17]

3.1.5. Según la documentación que acompaña a los textos

Otra clasificación que se puede hacer de los corpus es en función de si los textos que los componen están documentados o no.

1. *Corpus documentado*.- Corpus en el que cada documento que lo compone lleva asociado un archivo DTD (*Document Type Definition*) o una cabecera "*header*" de descripción de su filiación y sus constituyentes.[18]
2. *Corpus no documentado*.- Corpus en el que sus textos constituyentes no disponen de ningún apartado o archivo relacionado donde se describan sus elementos o su filiación.

3.2. CRITERIOS ESPECÍFICOS PARA LA CLASIFICACIÓN DE LOS CORPUS ORALES

En el apartado anterior se han definido una serie de criterios generales que permiten establecer distinciones genéricas entre diferentes tipos de corpus. Sin embargo, la especificidad de los corpus diseñados con vistas al análisis fonético o a las aplicaciones a las tecnologías del habla requiere establecer ciertos matices, que se abordan a continuación[19].

Podríamos considerar tres tipos de corpus: los orientados a la descripción fonética de la lengua (3.2.1), los que se utilizan para el desarrollo de sistemas en el ámbito de las tecnologías del habla (3.2.2) y los que propiamente se conocen como corpus orales, consistentes en transcripciones ortográficas de la lengua hablada (3.2.3.)

[16] Sinclair (1996:8) opina que las etiquetas estructurales no son suficientemente importantes como para considerar que un corpus es anotado si los textos que lo componen solo llevan este tipo de etiquetado.

[17] En este sentido debemos recomendar el uso del sistema propuesto por las llamadas "Normas TEI". *Guidelines for Electronic Text Encoding and Interchange (TEI P3)* presentadas en Sperberg-McQueen

y Burnard (eds.) (1994). Sobre la TEI, véase también Burnard (1995a) y Ide y Véronis (eds.) (1995). Puede encontrarse más información en las siguientes URLs: <http://www-tei.uic.edu/orgs/tei/> (*Text Encoding Initiative Home Page*), <http://etext.virginia.edu/TEI.html> (TEI *Guidelines for Electronic Text Encoding and Interchange P3*). Véase tambien el capítulo de Gerardo ARRARTE en este mismo libro sobre "Normas y estándares para la codificación de textos y para la ingeniería lingüística".

[18] Para información sobre DTD ver Sperberg-McQueen y Burnard (eds.) (1994) Para información sobre cabeceras (*header*) ver Ide (coord.) (1996).

[19] Para una presentación general de los corpus y bases de datos orales en el ámbito de la fonética y las tecnologías del habla véase Carré (1992), Lamel y Cole (1995) y Llisterri (1996c).

3.2.1. Corpus para la descripción fonética de la lengua

Aunque no constituyan exactamente corpus en el sentido en que aquí los estamos definiendo, cabe considerar en este apartado los inventarios de sistemas fonéticos y fonológicos de la lenguas del mundo utilizados en el estudio de los universales, integrados en bases de datos que permiten el análisis estadístico de la frecuencia de aparición de unidades segmentales o de rasgos fonéticos.

Sin embargo, los corpus para la descripción fonética de la lengua consisten tradicionalmente en materiales grabados en condiciones acústicas óptimas que permitan su posterior análisis experimental en el laboratorio. En estos casos solemos encontrar desde combinaciones de segmentos hasta fragmentos de habla espontánea, pasando por frases aisladas o por textos leídos. Lo que caracteriza a este tipo de corpus es un cuidadoso diseño del contenido, basado en el inventario de elementos segmentales y suprasegmentales de la lengua y un tamaño relativamente reducido, debido a que no suelen realizarse grabaciones con un número muy elevado de hablantes. Aún así, cada vez es mayor la tendencia a incluir producciones espontáneas y a utilizar grabaciones procedentes de los medios de comunicación por la diversidad de registos que pueden conseguirse y la relativa facilidad de obtención.

Es posible también diseñar un corpus con materiales equivalentes para varias lenguas, en el sentido de los corpus paralelos o comparables definidos en el apartado 3.1.1, con lo cual es posibles realizar estudios experimentales de fonética contrastiva. Estos mismos materiales pueden ser grabados por hablantes no nativos, con objeto de determinar los mecanismos de interferencia fonética que operan en la adquisición de segundas lenguas, o por hablantes con patologías del habla a efectos de analizar las desviaciones con respecto a los hablantes que no presentan estos problemas.

3.2.2. Corpus para el desarrollo de sistemas en el ámbito de las tecnologías del habla

El desarrollo y la validación de los sistemas de síntesis, reconocimiento y diálogo que han surgido en el campo conocido como las tecnologías del habla ha hecho necesario la constitución de corpus de naturaleza muy específica. En el caso de la creación de sistemas de conversión de texto a habla, es preciso disponer tanto de inventarios grabados de unidades de síntesis a partir de los cuales se realiza el paso de una representación ortográfica a una onda sonora, como de corpus que permitan el análisis de los elementos suprasegmentales para dotar al conversor de un modelo prosódico. Los sistemas de reconocimiento de habla

requieren también corpus grabados con las unidades fonéticas que se utilizarás en el reconocimiento, y en algunos casos corpus con materiales específicos como por ejemplo números de teléfono o de tarjetas de crédito orientados a determinadas aplicaciones del reconocimiento a lo servicios telefónicos automáticos. Ambas tecnologías necesitan también disponer de corpus textuales, a ser posible transcripciones de lengua oral lingüísticamente anotadas, para establecer los modelos probabilísticos de aparición de palabras sobre los que se basa el tratamiento lingüístico efectuado tanto en la síntesis como en el reconocimiento.

Un caso particular lo constituyen los corpus de diálogo utilizados para desarrollar y entrenar sistemas de interacción entre personas y máquinas, enfocados a ofrecer servicios automáticos a través del teléfono como la venta de billetes, la consulta de horrios de transportes públicos o los servicios bancarios. En este caso, suelen utilizarse corpus grabados y transcritos obtenidos mediante interacciones entre personas reales y una simulación del sistema de diálogo que se está construyendo, aunque también es útil el análisis de los diálogos naturales obtenidos en las situaciones comunicativas que se pretende modelar. En el caso de sistemas que incorporan además la traducción automática del habla, es imprescindible disponer de corpus orales paralelos en dos lenguas.

3.2.3. Transcripciones ortográficas de lengua hablada

En la lingüística de corpus tradicional se ha trabajado habitualmente con transcripciones ortográficas de la lengua hablada, procedentes de entrevistas realizadas especialmente para el corpus, de conversaciones espontáneas o de los medios de comunicación, incluyendose también otros materiales propios del registro oral como discursos políticos, clases, sermones, etc. Aunque el punto de partida sea una grabación, una vez transcrito, el corpus se trata con los mismos procedimientos que un corpus textual, enmarcándose plenamente en las caracterizaciones definidas en el apartado 3.1.

4. Principales aspectos en el diseño de un corpus

Una vez delimitados los distintos tipos de corpus y sus aplicaciones, es el momento de entrar en la discusión de los principales aspectos que deben considerarse en el diseño de un corpus[20]. Al igual que en los apartados anteriores, nos centraremos primero en las cuestiones generales, para introducir después aquellas que son específicas de algunas áreas de aplicación de los corpus.

4.1. Aspectos generales

4.1.1. Finalidad

El primer aspecto que hay que definir cuando se empieza a diseñar un corpus es la finalidad concreta para la que tiene que servir, aunque, como ya se ha dicho, se deba procurar que los

recursos lingüísticos sean siempre reutilizables. Este punto va a condicionar todos los demás, ya que es el que servirá de base para tomar las decisiones en todos ellos.

[20] Sobre el diseño de corpus véase, por ejemplo, Atkins *et al.* (1992), Leitner (1992), y Alvar y Corpas (1994) para el español.

4.1.2. Límites del corpus

Una vez especificada la finalidad, se han de establecer bien claramente los límites temporales, geográficos y/o lingüísticos que el corpus va a tener. Para ello se deberá marcar una fecha de inicio y otra de final y aclarar si las fechas se van a referir a la de los documentos originales o a la de las posibles copias transmisoras. Asimismo es necesario definir las lenguas que el corpus va a incluir y/o el área geográfica que abarcará.

Los límites temporales están muy condicionados al hecho de si el corpus es diacrónico o no. Pero incluso en el caso de los corpus sincrónicos[21] estos límites pueden variar substancialmente. En el corpus del español que se está recopilando en el King's College de Londres se recogen textos posteriores a 1990, mientras que en el corpus del español realizado por Alvar y otros en Biblograf se recogen textos publicados a partir de 1950. El *Longman Lancaster English Language Corpus* recoge textos posteriores al año 1899, y el *Corpus Textual Informatitzat de la Llengua Catalana* del *Institut d'Estudis Catalans* empieza la recolección de textos a partir de 1833, como fecha simbólica del inicio de la época moderna en cuanto al uso literario de la lengua.

Los límites geográficos también pueden variar mucho entre un corpus y otro; y no solamente los límites geográficos, sino también las distintas zonas territoriales que se marcan y los porcentajes de textos o palabras que se toman de cada zona. Para el español, por ejemplo, el corpus del King's College recoge un 25% de español de la Península, un 25% de español de Argentina y un 50% del español de las otras zonas de América del Sur, mientras que el corpus de Biblograf recoge el 60% de español peninsular, el 30% de español de América del Sur y el 10% de español de otras zonas.

Para el inglés, en el caso del corpus realizado en Birmingham dentro del proyecto COBUILD se ha establecido que se va a recoger un 70% de inglés de las Islas Británicas, un 20% de inglés de Estados Unidos y un 5% de inglés de otras partes (Sinclair (ed.), 1987). En cambio, el *Longman Lancaster English Language Corpus* ha establecido que el 50% de inglés será de las Islas Británicas, el 40% de inglés de Estados Unidos y el 10% restante de inglés de otras áreas geográficas.

4.1.3. Tipo de corpus

Una vez establecidos la finalidad y los límites hay que determinar el tipo de corpus que se va a realizar. Para ello será necesario definir cada uno de los parámetros siguientes: a) el porcentaje y la distribución de los diferentes tipos de textos que lo componen; b) la especificidad de los textos; c) la cantidad de texto que se tome de cada documento para

formar las muestras; d) la codificación y las anotaciones que se le añaden; e) la documentación que le acompañe.

Cada uno de estos puntos se ha tratado ya en el apartado anterior, pero la elección más controvertida es la referente a la cantidad de texto que se debe tomar de cada documento

[21] Aunque cualquir recopilación de textos tiene que ser obligatoriamente diacrónica porque casi nunca dos textos se han escrito en el mismo momento, cuando hablamos de corpus sincrónicos nos referimos a los que recogen muestras de la lengua de nuestro siglo.

Este punto ha sido bastante discutido y está íntimamente ligado a las posibilidades económicas, temporales y físicas (*hardware*) que tenga cada proyecto. Los corpus actualmente en preparación o los ya existentes adoptan diversos soluciones. Para el *Corpus del Castellano Contemporáneo* que se está preparando en el *King's College* de Londres, bajo la dirección del profesor Ife, la extensión media que se toma de cada texto es de 70.000 palabras. El *Longman/Lancaster English Language Corpus* incluye fragmentos de textos de unas 40.000 palabras, ya que su interés principal es el de "tener muchas fuentes diferentes más que textos completos". Por contra, el *International Corpus of English*, que dirige el profesor Greenbaum, solo recoge de cada documento fragmentos de 2.000 palabras, siguiendo el ejemplo del *Brown Corpus* y el *LOB Corpus* (Lancaster Oslo / Bergen). Por otro lado, también tenemos bastantes casos de proyectos que han decidido confeccionar el corpus solo con textos enteros; el ejemplo más conocido es el del COBUILD, actualmente con más de 20.000.000 de palabras.

John Sinclair, director del COBUILD, sintetitza su posición respeto a la conveniencia de trabajar con corpus de un tipo o de otro asumiendo que reuniendo textos enteros se evitan los problemas de las posibles diferencias que pueden haber entre distintas partes de un mismo texto, evitando así los inconvenientes de la validación de las muestras. Además, continua Sinclair, si es necesario, siempre es posible extraer muestras de una determinada longitud si se dispone de un corpus que recoja textos enteros. A corto plazo, el inconveniente de querer reunir un corpus textual es que con el mismo esfuerzo la cobertura de diferentes tipos de textos no será tan completa como la que puede proporcionar una colección de pequeñas muestras. Pero, a largo plazo, las ventajas de disponer de textos enteros son mayores.

Desde un punto de vista parecido, M. Alvar Ezquerra y sus colaboradores en el proyecto NERC (*Network of European Reference Corpora*) (Alvar y Villena (Coord.), 1994) , recomendaron la inclusión de textos enteros para el corpus del español, ya que consideraban que con los 20 millones de palabras propuestos como objetivo se podía abarcar un número importante de diferentes tipos de texto.

La inclusión de textos enteros en un corpus lo convierte en más abierto y apto para el estudio de un amplio abanico de aspectos lingüísticos. Además, siempre es más fácil recortar un texto entero que añadir fragmentos a los textos para completarlos.

Por otro lado, para obtener un corpus equilibrado es más fácil si se trabaja con corpus de referencia, sobre todo a corto plazo. Según Pierre Guiraud, una compilación de 300.000 palabras no ofrece garantías de ser equilibrada si las muestras son mayores de 500 palabras porque entonces aparecen pocas muestras (unas 600); tampoco la ofrece una compilación de

5 millones de palabras si las muestras se hacen más grandes de 2 o 3 mil palabras (unas 2.000).

Hay también quien opina que los corpus de referencia son poco adecuados para investigaciones estilísticas, pragmáticas, etc. porque las características discursivas de un texto se pierden cuando sólo disponemos de pequeñas partes. Las palabras y, sobre todo, las unidades fraseológicas necesitan ser examinadas dentro de la totalidad del discurso para poder comprender sus matices semánticos y pragmáticos. Pero este es un argumento más bien en contra de los corpus léxicos porque las muestras de los corpus de referencia suelen ser lo suficientemente largas como para que cada una contenga todo el sentido de las palabras o de las frases. En el caso de los corpus léxicos lo que interesa es el funcionamiento de las unidades léxicas dentro de las frases, pero no dentro del discurso. Sinclair opina que este tipo de corpus, por el hecho de estar compuesto por fragmentos muy escogidos y todos de la misma longitud, más que aportar imparcialidad lo que hace es dar una falsa idea de la realidad que quiere representar.

4.1.4. Proporciones de los diferentes grupos temáticos del corpus

Este es un punto bastante difícil de definir ya que las posibilidades pueden ser muchas y no hay unos criterios objetivos a los que podamos recurrir. De todos modos, es obvio que la definición de los diversos tipos y de las proporciones que se deben atribuir a cada uno de ellos es una cuestión en la que los sociólogos culturales deben tener mucho que decir. En los corpus *Brown* y *LOB,* por ejemplo, los textos están repartidos en 15 géneros, con una pequeña selección de textos elegida al azar en cada uno de ellos. El *Longman Lancaster English Language Corpus* está basado en muestras teóricas escogidas sin seguir ningún método estadístico. En este corpus se estableció recoger un 60% de textos informativos y un 40% de textos de creación, proporción extraida de las estadísticas de los libros más leídos en las bibliotecas. Las proporciones dentro de los textos escritos se establecieron en el 80% de libros, el 13,3% de periódicos y el 6,7% de otros medios. Dentro de estos porcentajes se establecieron, siguiendo el mismo sistema de obras más leídas, 10 grupos temáticos:

1	Ciencias purasy naturales	6,0%
2	Ciencias aplicadas	4,3%
3	Ciencias sociales	14,1%
4	Cuestiones mundiales	10,4%
5	Comercio y finanzas	4,4%
6	Artes	7,9%
7	Creencias y pensamientos	4,7%
8	Pasatiempos	5,7%
9	Ficción	40,0%
10	Poesía, teatro y humor	2,3%

Los distintos textos de cada grupo se seleccionaron utilizando el “Whitaker’s Books in Print”. Se dejaron de lado las traducciones, los textos no escritos totalmente en lengua inglesa, diccionarios y obras de referencia, trabajos de menos de 64 páginas, libros destinados

a niños de menos de 11 años, obras publicadas en países de habla no inglesa y trabajos en los que más del 75% del texto no era alfabético.

Una de las distribuciones más complejas pero a la vez más justificada es la que se hizo para el corpus de Birmingham, la cual no detallamos aquí por cuestiones de espacio.[22]

El corpus del español de Biblograf está distribuido en los siguientes grupos y proporciones:

1. no-ficción 25%
2. ficción 35%
3. periódicos 25%
4. panfletos 2,5%
5. cartas 2,5%
6. otros 10%

Por su parte, el corpus de español del King's College ha basado su criterio de selección en la última edición de la "*Dewey classifications*", clasificación utilizada en la mayoría de bibliotecas de todo el mundo. Para los libros, la selección principal se hizo a partir de los más vendidos, de los más recomendados en las universidades y de los sugeridos por expertos de cada tema.

4.1.5. Población y muestra

Como la finalidad de los corpus es la de describir el funcionamiento de la lengua a partir de una selección de textos lingüísticos, en el momento de construir uno es necesario aplicar los principios estadísticos de obtención de "muestras" representativas de una "población" [23]. Desafortunadamente, en algunas ocasiones es difícil poder aplicar las fórmulas de extracción de muestras porque es muy complejo (a veces imposible) delimitar el total de la población y además, en el caso de que ésta pueda ser delimitada, siempre habrá alguna característica de la población que no se habrá tenido en cuenta o no estará representada adecuadamente por las muestras. Otro factor que dificulta el muestreo en los corpus es el hecho de que no haya una unidad de la lengua evidente que se pueda usar para definir la población y las muestras, sino que a veces la unidad lingüística puede ser la palabra, otras veces la frase, otras el texto, etc.

Asimismo, todas las muestras son, de algún modo, tendenciosas. Los usuarios de los corpus tienen que estar evaluando continuamente los resultados obtenidos y, a la vista de ellos, ir corrigiendo las muestras. En todo momento, los investigadores se tienen que cuestionar cómo fueron obtenidas las muestras y hasta qué punto pueden ser válidas las conclusiones que de ellas se han extraído.

Un corpus siempre está construido a base de muestras con la intención de que de su observación se puedan extraer generalizaciones sobre la lengua; por eso, la relación entre

[22]

Esta distribución se puede encontrar en "Appendix 1: An Analysis of the Written Data in the Birmingham Main and Reserve Corpora" en Sinclair (ed.) (1987).

[23] La cuestión de la representatividad en el diseño de un corpus se trata, por ejemplo, en Biber (1993), Clear (1992) o de Haan (1992). En estos trabajos se abordan también algunas de las cuestiones discutidas en el apartado 4.1.4.

las muestras y la población es tan importante. De todos modos, la recopilación de una muestra representativa del total de la lengua es imposible. En el caso de los corpus que quieran representar la lengua general, la primera decisión que hay que tomar es la de si la muestra se va a escoger del lenguaje que se oye y lee (lenguaje de recepción: pocos productores pero muchos receptores), del lenguaje que se habla y escribe (lenguaje de producción: muchos productores con pocos receptores) o de ambos.

Cuanto más alto sea el grado de especialización de los diferentes grupos de la muestra más pequeños serán los problemas para seleccionar los textos que se deben incluir en cada uno de ellos.

El "constructor" de un corpus tiene que estar siempre muy atento a los aspectos de producción y recepción de los textos y, a pesar de que los textos de mucha recepción como los artículos periodísticos, son de fácil obtención, si se quiere que el corpus sea un reflejo real del uso de la lengua de los hablantes es necesario hacer todo lo posible para que también incluya textos de registros difíciles de obtener, como por ejemplo correspondencia personal. Definir la población en términos del lenguaje receptivo representa asignar mucho peso a una pequeña proporción de escritores y de hablantes cuyo *out put* de la lengua es recibido por una amplia audiencia a través de los medios de comunicación.

La producción puede estar muy influenciada por la recepción, pero solo la producción define la variedad de la lengua.

4.1.6. Número y longitud de los textos de la muestra

La selección de las partes de los textos de las que se van a extraer las muestras para un corpus de referenciase puede hacer de tres maneras: a) al azar; b) dividiendo los textos en tres partes de extensión parecida y extrayendo de cada una de ellas las muestras en número y proporciones aproximadamente iguales; c) determinando la estructura externa de los textos y decidiendo qué niveles estructurales se usarán para el muestreo (un número determinado de palabras o de frases de cada capítulo, un número determinado de cada apartado, un número determinado de cada párrafo, etc.).

Una vez establecidas las partes de los textos que se utilizarán para la extracción de las muestras, hay que acordar qué muestras se tomarán y la longitud que éstas deben tener.

Las muestras dentro de cada parte o sección definida se pueden seleccionar o bien escogiendo un número determinado de palabras o de oraciones a partir del inicio de cada sección, o bien haciendo una selección aleatoria entre las diferentes oraciones o los diferentes párrafos de cada sección. Normalmente se intenta que las muestras empiecen y terminen en un punto o en un punto y a parte.

Una vez definidas las secciones que se van a utilizar en cada texto para la extracción de las muestras, y establecido de dónde se tomaran las muestras dentro de cada sección, es necesario concretar el número de muestras y su longitud.

En el caso de los corpus de referencia, el número de palabras que se aconseja recoger de cada texto varía mucho según la finalidad y, sobre todo, las posibilidades tanto económicas como de equipamiento del proyecto. Se ha apuntado la conveniencia de recoger muestras de entre 2.000 y 70.000 palabras. De todos modos, los números y porcentajes que se han sugerido para la composición de muestras parecen bastante gratuitos, dado que ningún autor los ha justificado.

4.1.7. Captura de los textos y etiquetado

La introducción en el ordenador de los textos que tienen que configurar un corpus requiere tiempo, y el tiempo significa un coste considerable que puede condicionar el volumen que podrá tener el resultado final. Los textos impresos en papel pueden ser introducidos en el ordenador mediante un escáner y un programa informático de reconocimiento automático de caracteres (OCR: *Optical Character Recognition*). Con las mejoras que han experimentado últimamente los aparatos y los programas OCR—sobre todo al estar conectados a diccionarios de corrección--, la conversión de texto impreso a texto en formato electrónico está siendo cada vez más efectiva[24]. Alternativamente, el texto impreso también puede ser informatizado de forma manual tecleándolo directamente al ordenador, pero, está claro, que ésta debe ser la última opción, reservada solamente para las transcripciones de cintas o para la recuperación de textos impresos en muy mal estado o de formato complicado, dos casos en que el escáner ofrece pocas garantías y requiere mucho trabajo de revisión.

De todos modos, para dar por bueno un texto que se ha introducido en el ordenador mediante un escáner se recomienda:

1.- Escanear el texto dos veces y realizar un control del resultado por parte de dos personas distintas.
2.- Comparar automáticamente los dos ficheros, comprobando con el original cada punto de divergencia.
3.- Realizar una lista de frecuencias para revisar sobre todo las unidades de una sola aparición (no es normal cometer varias veces el mismo error).
4.- Efectuar una lectura de la última versión entre dos personas trabajando juntas.

A veces los textos se pueden obtener directamente en formato electrónico, ya sea porque otra persona los había introducido para un uso propio ya sea porque originariamente se habían hecho en este formato. Actualmente, a través de Internet se puede acceder a gran cantidad de textos digitalizados de todo tipo. Para la confección de corpus textuales son especialmente interesantes los periódicos y publicaciones a los que esta red da acceso. Este sistema de captura de textos elimina costes y posibilidades de errores, siendo solamente necesario adaptar los archivos importados a los formatos usados en el corpus.

Los textos ya digitalizados que forman un corpus deberán ser marcados con determinados códigos - codificación - para señalar sus elementos estructurales, para especificar las características de sus fuentes originales, para marcar determinadas informaciones importantes para su explotación, etc. El tema de la codificación y etiquetado

[24] Véase, por ejemplo, Belaïd (1995).

de los textos es importantísimo para facilitar la posterior explotación del corpus y, por lo tanto, este aspecto tiene un peso considerable en la planificación y en los costes de cada proyecto.

Precisamente el alto coste que supone la codificación y el etiquetado de los textos (ya sea en términos de tiempo o en términos económicos) ha impulsado la idea de definir estándares de codificación y etiquetado para facilitar el intercambio y la reusabilidad de los textos ya preparados. Actualmente hay un consenso creciente en que las marcas SGML (*Standard Generalized Markup Language*) proveen una base adecuada para un esquema estándar y que la TEI (*Text Encoding Initiative*), basada precisamente en este sistema[25] proporciona un buen procedimiento para la codificación de textos en formato electrónico; las propuestas desarrolladas por Ide (Coord.) (1996) en el marco de los proyectos
EAGLES *(Expert Advisory Group on Language Engineering Standards)* y MULTEXT (*Multilingual Text Tools and Corpora*) constituyen, sin duda, una aportación importante en el ámbito de la codificación de corpus. En lo que se refiere a la anotación lingüística mediante etiquetas que definan, por ejemplo, partes de la oración, existe una mayor diversidad de sistemas, entre los que cabe destacar las recogidas en las *Guidelines* de EAGLES[1]. Esta anotación puede, naturalmente, llevarse a cabo utilizando los mecanismos propios del SGML.

4.1.8. Procesamiento del corpus

El corpus por sí solo no es suficiente para facilitar datos exhaustivos del comportamiento del lenguaje. Para poder aprovechar al máximo las informaciones que contiene es necesario poder disponer de herramientas adecuadas para su procesamiento y para su explotación. En este sentido hay que decir que tan importante es el corpus como las herramientas. Actualmente se trabaja en programas de gran complejidad destinados a la lingüística de

[1] En el marco del proyecto EAGLES se ha propuesto un esquema para la anotación morfosintáctica de textos (Leech y Wilson, 1996) y unas orientaciones preliminares para la anotación sintáctica (Leech *et al*., 1996). En lo que se refiere al español, puede verse, por ejemplo, una propuesta de codificación en SGML de la anotación morfosintáctica desarrollada para el Corpus de Referencia del Español Actual (CREA) de la Real Academia Española en Pino y Santalla (1996).

[27] El apéndice B de McEnery y Wilson (1996) ofrece información sobre estas herramientas, así como el *Natural Language Software Registry*, que puede consultarse en <http://cl-www.dfki.unisb.de/cl/registry/>. Algunas muestras de herramientas desarrolladas para el español se describen en los diversos trabajos publicados en en *Procesamiento del Lenguaje Natural*, revista de la Sociedad Española para el Procesamiento del Lenguaje Natural <http://gplsi.ua.es/sepln/>.

corpus, así que ya se dispone de un buen número de ellos destinados a tareas muy específicas[27]. Entre los trabajos básicos que deben facilitar los programas para explotación de corpus en el campo de la lingüística cabe destacar:

- frecuencia de aparición de palabras
- índices y concordancias
- lematización

[25] Sperberg-McQueen y Burnard, (eds.) (1994).

- análisis morfológico (*tagging*)
- análisis sintáctico (*parsing*)
- desambiguación semántica
- detección de unidades recurrentes (*collocations*)

4.1.9. Crecimiento del corpus y "Feedback"

Con la finalidad de tener un corpus equilibrado es conveniente adoptar un método de aproximaciones sucesivas. Primero, en su preparación hay que procurar conseguir un corpus representativo; después, al utilizarlo, hay que analizar los resultados y detectar sus puntos débiles respecto de la representatividad. A la vista de estos análisis, se debe ir reajustando las proporciones del corpus constantemente. Para ello es necesario colaborar conjuntamente con expertos en estadística que aporten métodos para mejorar el equilibrio del corpus y estar en constante contacto con los usuarios ya que ellos son los que mejor detectaran sus limitaciones.

4.1.10. "Hardware" y "software"

Un aspecto también muy importante que hay que tratar al diseñar un corpus es el de la estimación de la infraestructura informática, tanto en su componente de *hardware* (aparatos) como en el de *software* (programas), que se va a necesitar para poder desarrollarlo y explotarlo. Las necesidades de infraestructura dependerán, como es lógico, de la extensión que deba tener el corpus, de los diferentes procesos que se deban realizar y de la naturaleza oral o textual de los materiales. Almacenar simplemente los textos de un corpus es una tarea que necesita poco equipamiento y escasos programas, pero tenerlo dispuesto para una fácil recuperación de la información y para la realización de procesos de análisis requiere ya ordenadores preparados (generalmente estaciones de trabajo) y programas sofisticados, en algunos casos realizados *ad hoc*.

4.1.11. Aspectos legales

Uno de los problemas más difíciles de resolver, principalmente por su carácter no filológico ni científico, es el de los derechos de autor (*copyright*). Esta cuestión se convierte en

trascendental cuando se trata de corpus que usan fuentes literarias o periodísticas y al que se quiere dar difusión para su explotación. El problema se hace más difícil por el hecho de que en muchos casos la legislación no ofrece soluciones claras; hay algunos países, por ejemplo, que tienen un consenso para conceder ciertos privilegios a las universidades. Tampoco está bien definida la normativa a que está sujeta la reproducción y utilización de los textos periodísticos capturados a través de Internet, o el límite de palabras seguidas que se pueden copiar para no incumplir la normativa de los derechos de autor.

Es necesario y justo proteger, mediante el *copyright*, los derechos de los autores y de las editoriales sobre los textos que ellos han creado o publicado. Es necesario revisar y ampliar la normativa actual como respuesta al rápido desarrollo de las técnicas informáticas de captura de textos. Es probable que cualquier texto editado (o parte considerable de texto) que tenga que ser computerizado e incluido en un corpus esté bajo esta ley y se necesite pedir autorización para su uso.

Las siguientes consideraciones son importantes al tratar de los derechos de autor y el corpus:

- ¿El texto está protegido por la ley de los derechos de autor? La legislación varía según los países pero por norma general la duración de los derechos es limitada.
- La transcripción de textos orales registrados de un medio de comunicación (radio, televisión) también está sujeta a esta normativa.
- La difusión de grabaciones que no proceden de los medios de comunicación requiere el permiso escrito de los hablantes, obtenido en general con posterioridad a la realización de las mismas para no restar espontaneidad al intercambio comunicativo. Es necesario también proteger la intimidad de las personas, cambiando, por ejemplo, sus nombres por iniciales.
- Aunque se paguen pequeñas cantidades por cada texto incluido en un corpus, si el corpus es grande, los trabajos administrativos y el total que se debe pagar pueden ser considerables, de manera que solo algunas organizaciones con importantes medios que se aseguren su explotación podrán justificar los costes.
- En el caso de la cesión desinteresada de los derechos, los propietarios de los derechos de autor tienen que tener la seguridad de que la compilación del corpus no será inconveniente para el potencial de ganancias y de que no habrá ninguna explotación comercial directa del corpus.
- La posible explotación y distribución de un corpus tiene que estar cuidadosamente pactada con los propietarios de los derechos de autor de los textos que lo componen.
- Si el corpus se ha hecho con finalidades comerciales, tienen que constar los propietarios de los derechos de autor.

4.1.12. Presupuesto y etapas

Una vez definidas todas las cuestiones mencionadas hasta este momento sólo hace falta establecer las diferentes etapas en que se va a realizar el proyecto y cómo se va a llevar a cabo su mantenimiento (en el caso de tratarse de un corpus abierto). Ello implica la realización de un presupuesto teniendo en cuenta tanto los costes del personal humano como los de los programas y ordenadores y demás aparatos, así como los de la adquisición de los derechos de autor en el caso de que los textos utilizados así lo requieran.

4.2. ASPECTOS ESPECÍFICOS DE LOS CORPUS ORALES

El diseño y las distintas fases de elaboración de corpus orales que incluyen grabaciones de la señal sonora tiene algunos aspectos específicos que, complementando los más generales discutidos en el apartado anterior, se exponen a continuación. Es preciso tener en cuenta que en lo que se refiere especialmente a la creación de corpus para las aplicaciones propias de las tecnologías del habla se han desarrollado propuestas de estandarización para cada una de las fases de la constitución de un corpus en el marco de los proyectos europeos SAM (*Speech Assessment Methodologies*) y EAGLES (*Expert Advisory Group on Language Engineering Standards*), que actualmente constituyen una referencia esencial en el momento de abordar este tipo de corpus [28].

4.2.1. Adquisición de los datos

En los corpus orales a los que aludíamos en el apartado 3.2. la adquisición de los datos requiere necesariamente la realización de grabaciones o, alternativamente, su obtención a través de la radio y la televisión o de archivos sonoros que se encuentren disponibles. Si el objetivo del corpus es el análisis de la lengua oral (*cf.* 3.2.3), es suficiente con que la grabación tenga la calidad necesaria para permitir una transcripción ortográfica sin dificultades. En cambio, si pretendemos realizar un trabajo experimental en fonética (*cf.* 3.2.1.) o desarrollar los sistemas propios de las tecnologías del habla (*cf.* 3.2.2), el material sonoro debe reunir unas características específicas, para lo cual la grabación debe realizarse en un entorno acústico controlado como una cabina insonorizada o anecoica y por procedimientos digitales.

Mención aparte merecen los corpus para el estudio articulatorio del habla, que requieren técnicas más complejas para recoger los movimientos del aparato fonador; también debemos referirnos a los diversos métodos desarrollados para la obtención de producciones orales controladas que mantengan a la vez un cierto grado de espontaneidad, como por ejemplo la denominada "tarea del mapa" (Anderson *et al*., 1991), o que permitan el análisis fonético de los diversos estilos de habla (Péan *et al*., 1993). Igualmente constituyen un caso específico los corpus recogidos a través del teléfono a fin de entrenar y evaluar sistemas de reconocimiento de habla[29] y los que se recopilan mediante el procedimiento conocido como 'el Mago de Oz' a fin de modelar la interacción entre un sistema automático y un usuario (Fraser y Gilbert, 1991).

[28] Los principales resultados del proyecto SAM se recogen en Fourcin *et al.* (1989) y se resumen en Fourcin y Dolmazon (1991). Las recomendaciones elaboradas por EAGLES en el campo de los corpus orales se exponen en Gibbon *et al.* (1997); un resumen delos trabajos realizado en EAGLES se encuentra en Winski *et al.* (1995).

[29] Véanse, por ejemplo, los trabajos realizados en el marco de los proyectos SPEECHDAT (*Spoken Language Resources*) <http://www.icp.grenet.fr/SpeechDat/home.html>. y SPEECHDAT II (*Speech Databases for the Creation of Voice Driven Teleservices*) <http://www.phonetik.unimuenchen.de/SpeechDat.html>.

4.2.2. Selección de locutores

A los problemas de la delimitación del corpus y de la selección de las muestras discutidos en los apartados anteriores, se une, en el caso de los corpus orales, el de la selección de los locutores. Los criterios utilizados varían, naturalmente, en función de los objetivos del corpus, pero suelen incluir el sexo, la edad, la procedencia espacial y el nivel sociocultural, pudiéndo tenerse también en cuenta hábitos que pueden originar patologías vocales como el uso del tabaco. Mientras que en algunos estudios se pretende reflejar las características fonéticas de un grupo reducido de hablantes considerado representativo, en corpus diseñados para desarrollar servicios telefónicos que pueden ser utilizados por toda la población, suele emplearse una estrategia de recogida de datos que garantice la presencia de muestras procedentes de un gran número de locutores aunque las muestras de habla sean relativamente breves. La selección del locutor plantea, en estos casos, problemas análogos a la selección de textos en un corpus que pretenda una cobertura general.

4.2.3. Procesamiento del corpus

Al igual que en el caso de los corpus textuales, una vez se han recogido los materiales de base, debe llevarse a cabo un procesamiento de los mismos que permita su utilización posterior. El primer paso suele ser la transcripción ortográfica, que en determinado tipo de corpus se acompaña de una transcripción fonética o fonológica. A continuación, a cada segmento de la onda sonora se le asocia una etiqueta que lo define en términos fonéticos o fonológicos (*labelling*) y se lleva a cabo la alineación (*alignment*) entre la señal sonora y las etiquetas, obteniendo una representación que puede compararse a la de una partitura musical con la letra correspondiente. El proceso de etiquetado segmental puede llevarse a cabo a varios niveles (Barry y Fourcin, 1992; Tillmann y Pompino-Marschall, 1993) y completarse con una anotación de las características suprasegmentales, codificadas según diversos sistemas que se exponen en el siguiente apartado.

Si se cumplen todas las etapas, es posible llegar a disponer de un corpus que contenga la señal sonora sincronizada con la transcripción ortográfica y la transcripción fonética o fonológica, de modo que, una vez definida una estructura de base de datos, el corpus pueda ser consultado partiendo de etiquetas fonéticas, de marcas prosódicas o de la transcripción ortográfica al tiempo que se accede a la grabación correspondiente.

Los corpus de lengua oral que consisten únicamente en transcripciones ortográficas - ya que no suele ser factible realizar una transcripción fonética completa de un número elevado

de horas de grabación - conllevan un procesamiento menos complejo, aunque en algunos casos contienen marcas prosódicas útiles para el análisis del discurso o de la conversación.

4.2.4. La transcripción fonética segmental y suprasegmental

Como acabamos de ver, un corpus puede enriquecerse con anotación lingüística, que en el caso de los corpus orales para determinadas aplicaciones, suele ser de tipo fonético segmental o suprasegmental.

En lo que respecta a la transcripción fonética segmental, suele recomendarse el uso del Alfabeto Fonético Internacional (AFI)[30]. Sin embargo, las necesidades del intercambio electrónico de textos han llevado a establecer una codificación de los símbolos del AFI (Esling y Gaylor, 1993). En el campo de las tecnologías del habla en el contexto europeo es de uso común el sistema de transcripción conocido como SAMPA (SAM *Phonetic Alphabet*), que utiliza los símbolos presentes en un teclado convencional y ha sido adaptado a buena parte de las lenguas europeas; una propuesta más reciente - conocida como X-SAMPA - sugiere la ampliación del sistema para codificar los símbolos del AFI[31].

En lo que respecta a la transcripción de los elementos suprasegmentales, además del conjunto de símbolos del AFI, se dispone también de varios sistemas que pueden ser utilizados en corpus en soporte electrónico[32], sin que parezca existir, por el momento, unanimidad en cuanto a la utilización preferente de ninguno de ellos. Entre los más difundidos cabe citar ToBI (*Tone and Break Index*) (Silverman *et al*., 1991), SAMPROSA (*SAM Prosodic Alphabet)*, desarrollado en el marco del proyecto SAM (Gibbon, 1989)[2] e INTSINT (*International Transcription System for Intonation*) (Hirst *et al*., 1994)[34]. Además de los procedimientos mencionados, propios del ámbito de la fonética y las tecnologías del habla, desde la perspectiva de la lingüística de corpus se han desarrollado también sistemas de anotación prosódica de corpus orales ortográficamente transcritos en los que suelen marcarse pausas, unidades tonales, cambios de intensidad, de rango melódico o de velocidad de elocución - como en el caso de las convenciones de la TEI para la transcripción de corpus orales (Johansson 1995a,b) - o bien sílabas acentuadas, sílabas prominentes no acentuadas y movimientos tonales como en la propuesta de French (1992) para el proyecto NERC.

4.2.5. La transcripción y codificación de la lengua oral

La transcripción ortográfica de los corpus de lengua oral tal como se describen en 3.2.3. plantea diversos problemas entre los que se cuentan las variaciones en la pronunciación no recogidas en los diccionarios normativos, el uso de los signos de puntuación y la representación de siglas, abreviaturas, palabras deletreadas o secuencias numéricas.

---

[2] SAMPROSA se describe también en <http://www.phon.ucl.ac.uk/home/sampa/samprosa.htm>. [34] Puede obtenerse más información sobre INTSINT en <http://www.lpl.univaix.fr/~hirst/intsint.html>.

A estas cuestiones debe sumarse la representación de los elementos propios de la lengua oral como las pausas, la delimitación de los enunciados y de las unidades tonales, las variaciones en los elementos suprasegmentales, los elementos vocales tanto semi-léxicos (por ejemplo las denominadas 'pausas llenas') como no léxicos (por ejemplo risas o toses) , los cambios de turno de palabra, las intervenciones simultáneas de varios hablantes o las dudas, palabras truncadas, repeticiones y errores de producción corregidos o no por el propio hablante. Se trata aquí de fenómenos tradicionalmente tratados por los especialistas en análisis del discurso y de la conversación[35], cuya codificación representa un enriquecimiento de la transcripción ortográfica y resultad indispensable para determinadas utilizaciones de los corpus en el análisis lingüístico.

Proyectos de naturaleza tan diversa como NERC dedicado a definir corpus de refrencia o *SpeechDat* centrado en la creación de bases de datos para el desarrollo o la evaluación de sistemas de reconocimiento de habla han creado una serie de convenciones para la transcripción ortográfica - a veces denominada transliteración - de la lengua oral[36]. Los posibles estándares tanto en lo que se refiere a la transcripción ortográfica como a la codificación de la lengua oral se han abordado en el marco de la TEI y de EAGLES, intentando combinar las necesidades de la lingüística de corpus con las de las tecnologías del habla[37].

Sin embargo, al igual que en la transcripción de los elementos prosódicos, no parece que dispongamos aún de un procedimiento unánimente aceptado y utilizado en lo que se refiere a la codificación. Por ello, parece recomendable la utilización de sistemas que permitan una traducción (semi)automática entre diversas propuestas y que no sean incomptibles con procedimientos estandarizados como el etiquetado en SGML, facilitando además al máximo la labor de transcriptor mediante un sistema de ayuda a la codificación.

5. Conclusiones

En este capítulo se ha intentado presentar, por una parte, una definición de 'corpus' complentada por una tipología que permita deslindar los corpus en sentido estricto de otro tipo de recopilaciones de materiales lingüísticos y por una breve reseña de sus principales aplicaciones. Por otra parte, se ha intentado ofrecer también algunas indicaciones sobre las principales etapas propias del proceso de elaboración de un corpus, teniendo en cuenta tanto la lengua escrita como la hablada. Con ello se ha querido mostrar que nos encontramos ante

[30] La última revisión del AFI aparece en IPA (1993); en IPA (1995) puede encontrarse una versión preliminar del *IPA Handbook*, de próxima publicación.

[31] Véanse sobre SAMPA Wells (1989) y la información recogida en <http://www.phon.ucl.ac.uk/home/sampa/home.htm> donde aparecen las adaptaciones a diversas lenguas realizadas hasta el momento; X-SAMPA se presenta en Wells (1994) y en <http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm>.

[32] En Gibbon (1989) y en Llisterri (1994) se presenta una revisión de diversos sistemas de anotación prosódica.

una poderosa herramienta que puede dar lugar tanto a investigaciones sobre el uso o la evolución de la lengua como al desarrollo de productos en el marco de la denominada ingeniería lingüística.

[35] Véanse, por ejemplo, Du Bois (1991), Du Bois *et al*. (1993), Edwards (1993, 1995), Gumperz y Berenz (1993) o Payrató (1995). Una buena muestra de las prácticas de transcripción y codificación de textos orales en la lingüística de corpus se encuentra en la recopilación de Leech *et al.* (eds.) (1995).

[36] Las convenciones de transcripción ortográfica del proyecto NERC se encuentran en French (1992) y se desarrollan para el español en Villena (1994); las recomendaciones sobre la transcripción de SpeechDat se encuentran en Winski *et al.* (1996).

[37] Las propuestas de la TEI se dicuten en Johansson (1995a,b) y las de EAGLES se recogen en el capítulo dedicado a la representación textual del *EAGLES Handbook on Spoken Language Systems* y en Llisterri (1996b).

Es importante insistir aquí en que los corpus son recursos que pueden ser utilizados de muy diversas maneras; sin embargo, en la fase de diseño, es imprescindible disponer de una definición clara de los objetivos que guían la constitución del corpus, sin la cual resulta extraordinariamente difícil enfrentarse a las múltiples opciones metodológicas que se plantean en esta primera etapa. Un segundo aspecto esencial es que, dado el esfuerzo económico y humano que supone la creación de un corpus, parece lógico pensar en que éste debe poder ser reutilizado por otros investigadores y para fines diferentes a los que fue concebido. Para ello es del todo necesario intentar adaptarse al máximo a los estándares existentes o, cuando existen varias alternativas, considerar la posibilidad de una conversión relativamente poco costosa y lo más automatizada posible.

Por estos dos motivos, el diseño - tanto del contenido como del modo de representaciónse plantea como la etapa más importante en la constitución de un corpus o de cualquier recurso lingüístico. Invertir esfuerzos en esta actividad es, a nuestro modo de ver, la mejor manera de garantizar el éxito del proyecto y de facilitar el intercambio y la reutilización del producto final.

6. Referencias bibliográficas

AARTS, J. - de HAAN, P. - OOSTDIJK, N. (eds.) (1993) *English Language Corpora: Design, Analysis and Explotation*, Amsterdam: Rodopi.

AARTS, J.- MEIJS, W. (eds.) (1984) *Corpus Linguistics. Recent Developments in the Use of Corpora in English Language Research*, Amsterdam: Rodopi.

AARTS, J.- MEIJS, W. (eds.) (1986) *Corpus Linguistics II. New Studies in the Analysis and Exploitation of Computer Corpora*, Amsterdam: Rodopi.

AARTS, J.- MEIJS, W. (eds.) (1990) *Theory and Practice in Corpus Linguistics*, Amsterdam: Rodopi.

AIJMER, K.- ALTENBERG, B. (eds.) (1991) *English Corpus Linguistics. Sudies in Honour of Jan Svartvik*, London: Longman.

ALVAR EZQUERRA, M.- CORPAS PASTOR, G. (1994) "Criterios de diseño para la creación de córpora", in ALVAR EZQUERRA, M.- VILLENA PONSODA, J.A. (Coord) *Estudios para un corpus del español*, Málaga: Universidad de Málaga, pp. 31-40.

ALVAR EZQUERRA, M.- VILLENA PONSODA, J.A. (Coord.) (1994) *Estudios para un corpus del español,* Málaga: Universidad de Málaga (Anejo 7 de Analecta Malacitana, Revista de la Sección de Filología de la Facultad de Filosofía y Letras de Málaga).

ANDERSON, A.H. - BADER, M.- BARD, E.G.- BOYLE, E.- DOHERTY, G.- GARROD, S.- ISARD, S.KOWTKO, J.- McALLISTER, J.- MILLER, J.- SOTILLO, C.- THOMPSON, H.S.- WEINERT, R. (1991) " The HCRC Map Task corpus", *Language and Speech*, 34, 4, pp. 351-366.

ATKINS, S.- CLEAR, J.- OSTLER, N. (1992) "Corpus design criteria", *Literary and Linguistic Computing* 7, 1, pp. 1-16.

ATKINSON, D.- BIBER, D. (1994) "Register: A Review of Empirical Resarch", en BIBER, D.- FINEGAN, E. (eds) *Sociolinguistic Perspectives on Register,* Oxford - New York: Oxford University Press, pp. 351-385.

BADIA, T.- CABRÉ, M.T.- LLISTERRI, J.- DE YZAGUIRRE, Ll. (1994) *Recursos en llengua catalana: estat de la qüestió.* Jornada de Compatibilitat i accessibilitat dels corpus de dades en llengua catalana. Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

BARRY, W.J.- FOURCIN, A.J. (1992) "Levels of Labelling", *Computer Speech and Language*, 6, pp. 1-14.

BELAÏD, A. (1995) "OCR: Print", in COLE, R.A.- MARIANI, J.- USZKOREIT, H.- ZAENEN, A.- ZUE, V. (eds.) Survey of the State of the Art in Human Language Technology, pp. 81-85. Publicación electrónica en URL: <http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html>.

BIBER, D. (1990) " Methodological issues regarding corpus-based analyses of linguistic variation", *Literary and Linguistic Computing*, 5, 4, pp. 257-269.

BIBER, D. (1993) "Representativeness in corpus design", *Literary and Linguistic Computing*, 8, 4, pp. 243257.

BIBER, D.- FINEGAN, E. (1991) " On the exploitation of computerized corpora in variation studies", in AIJMER, K.- ALTENBERG, B. (eds) *English Corpus Linguistics. Sudies in Honour of Jan Svartvik*, London: Longman, pp. 204-220.

BLECUA, J. M. (en prensa) "Els corpus lingüístics orals", en CLUB-1, *1er Col.loqui Lingüístic de la Universitat de Barcelona,* Universitat de Barcelona, 20 de desembre de 1993. Secció de Lingüística Catalana, Departament de Filologia Catalana, Universitat de Barcelona.

BURNARD, L. (1995a) "The Text Encoding Initiative: an overview", en LEECH, G.- MYERS, G.THOMAS, J. (eds) *Spoken English on Computer: Transcription, Markup and Applications*, Harlow: Longman, pp. 69-81.

BURNARD, L. (1995b) "Text Encoding for Information Interchange. An Introduction to the Text Enxoding Initiative", TEI Document no. TEI J31, July 1995. Publicación electrónica en URL: <http://www.uic.edu/orgs/tei/info/teij31/>.

CARRÉ, R. (1992) "Speech Databases" en AINSWORTH, W.A. (Ed) *Advances in Speech, Hearing and Language Processing. A Research Annual, Volume 2*, London: Jai Press, pp. 199-216.

CLEAR, J. (1992) "Corpus sampling", en LEITNER, G. (Ed) *New Directions in English Language Corpora. Methodology, Results, Software Development*, Berlin: Mouton de Gruyter, pp. 21-32.

COLE, R. (ed.) (1996) "Language Resources", en COLE, R.A.- MARIANI, J.- USZKOREIT, H.- ZAENEN, A.- ZUE, V. (eds.) (1997) *Survey of the State of the Art in Human Language Technology*, Cambridge: Cambridge Univversity Press, pp. 441-474. Publicación electrónica en URL: <http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html>.

de HAAN, J. P. - OOSTDIJK, N. (eds.) (1993) *English Language Corpora: Design, Analysis and Exploitation*, Amsterdam: Rodopi

de HAAN, P. (1992) "The optimum corpus sample size?", en LEITNER, G. (Ed) *New Directions in English Language Corpora. Methodology, Results, Software Development*, Berlin: Mouton de Gruyter, pp. 3-20.

DU BOIS, J.W. (1991) " Transcription design principles for spoken discourse research", *Pragmatics* 1, pp. 71-106.

DU BOIS, J.W.- SCHUETZE-COBURN, S.-CUMMING, S.- PAOLINO, D. (1993) "Outline of discourse transcription", en EDWARDS, J.A.- LAMPERT, M.D. (eds.) *Talking Data: Transcription and Coding in Discourse Research,* Hillsdale, N.J.: Lawrence Erlbaum Associates, pp. 45-90.

EDWARDS, J.A. (1993) "Principles and Contrasting Systems of Discourse Transcription", en EDWARDS, J.A.- LAMPERT, M.D. (eds.) *Talking Data: Transcription and Coding in Discourse Research*, Hillsdale, N.J.: Lawrence Erlbaum Associates, pp. 3-32.

EDWARDS, J.A. (1993) "Survey of Electronic Corpora and Related Resources for Language Researchers", en EDWARDS, J.A.- LAMPERT, M.D. (eds.) *Talking Data: Transcription and Coding in Discourse Research*, Hillsdale, N.J.: Lawrence Erlbaum Associates, pp. 263-310.

EDWARDS, J.A. (1995) " Principles and alternative systems in the transcription, coding and mark-up of spoken discourse", en LEECH, G.- MYERS, G.- THOMAS, J. (eds.) *Spoken English on Computer: Transcription, Markup and Applications*. Harlow: Longman, pp. 19-34.

ESLING, J.H.- GAYLORD, H. (1993) "Computer Codes for Phonetic Symbols", *Journal of the International Phonetic Association*, 23, 2, pp. 77-82.

FERNÁNDEZ, A.- LLISTERRI, J. (1996) *Informe sobre recursos lingüísticos para el español (II): Corpus escritos y orales disponibles y en desarrollo en España*, Alcalá de Henares: Observatorio Español de Industrias de la lengua, Instituto Cervantes.

FOURCIN, A.- DOLMAZON, J.M. (on behalf of the SAM Project) (1991) "Speech knowledge, standards and assessment" en *Actes du XIIème Congrès International des Sciences Phonétiques. 19-24 août 1991, Aixen-Provence, France*, Aix-en-Provence: Université de Provence, Service des Publications, Vol 5, pp. 430433.

FOURCIN, A.- HARLAND, G.- BARRY, W. - HAZAN, V (eds.) (1989) *Speech Input and Output Assessment. Multilingual Methods and Standards,* Chichester: Ellis Horwood Ltd.

FRASER, N.- GILBERT, G.N. (1991) "Simulating speech systems", *Computer Speech and Language*, 5, 1, pp. 81-99.

FRENCH, J.P. (1992) *Transcription proposals: multilevel system*, Working paper, University of Birmingham, October 1992. NERC-WP4-50.

FRIES, U.- TOTTIE, G.- SCHNEIDER, P. (eds.) (1994) *Creating and Using English Language Corpora*, Amsterdam: Rodopi.

GIBBON, D. (1989) "Survey of Prosodic Labelling for EC Languages". SAM-UBI-1/90, 12 February 1989; Report e.6, en ESPRIT 2589 (SAM) *Interim Report, Year 1.* Ref. SAM-UCL G002, University College London, February 1990.

GIBBON, D. - MOORE, R. - WINSKI, R. (eds.) (1997) *Handbook of Standards and Resources of Spoken Language Systems*, Berlin: Mouton de Gruyter. Publicación electrónica en URL: <www.degruyter.de/EAGLES/>.

GUMPERZ, J.J.- BERENZ, N. (1993) " Transcribing Conversational Exchanges", en EDWARDS, J.A.LAMPERT, M.D. (eds.) *Talking Data: Transcription and Coding in Discourse Research*, Hillsdale, N.J.: Lawrence Erlbaum Associates, pp. 91-122.

HIRST, D.J.- IDE, N. - VÉRONIS, J. (1994) "Coding fundamental frequency patterns for multi-lingual synthesis with INTSINT in the MULTEXT project", en *Conference Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*. September 12-15, 1994, Mohonk Mountain House, New Paltz, New York, USA, pp. 77-80.

IDE, N. (Coord.) (1996) *Corpus Encoding Standard*. Document CES 1. Version 1.4. October, 1996. Publicación electrónica en URL: <http://www.cs.vassar.edu/CES/>.

IDE, N.- VÉRONIS, J. (eds.) (1995) The Text Encoding Initiative: Background and Contexts. *Computers and the Humanities* 29, 1-3; publicado en forma de libro en Dordrecht: Kluwer Academic Publishers.

IPA (1993) "IPA Chart, revised to 1993", *Journal of the International Phonetic Association* 23,1. Publicación electrónica en URL: <http://www.arts.gla.ac.uk/IPA/ipachart.html>.

IPA (1995) Preview of the IPA Handbook, *Journal of the International Phonetic Association*, 25, 1.

JOHANSSON, S. (1995a) "The Encoding of Spoken Texts", *Computers and the Humanities* 29, 1, pp. 149158; en IDE, N.- VÉRONIS, J. (eds) (1995) *The Text Encoding Initiative. Background and Context*, Dordrecht: Kluwer Academic Publishers, pp. 149-158.

JOHANSSON, S. (1995b) “The approach of the Text Encoding Initiative to the encoding of spoken discourse”, en LEECH, G.- MYERS, G.- THOMAS, J. (eds.) *Spoken English on Computer: Transcription, Markup and Applications*, Harlow: Longman, pp. 82-98.

JOHANSSON, S. (ed.) (1982) *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities.

JOHANSSON, S.- STENSTRÖM, A. (eds.) (1991) *English Computer Corpora: Selected Papers and Research Guide*, Berlin: Mouton de Gruyer (Topics in English Linguistics, 3).

KNWOLES, G. (1990) “The use of spoken and written corpora in the teaching of language and linguistics”, *Literary & Linguistic Computing*, 5, 1, pp. 45-48.

KYTÖ, M. - IHALAINEN, O.- RISSANEN, M. (eds.) (1988) *Corpus Linguistics, Hard and Soft. Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora*, Amsterdam: Rodopi.

KYTÖ, M.- RISSANEN, M.- WRIGHT, S. (eds.) (1994) *Corpora across the Centuries*, Amsterdam: Rodopi.

LAMEL, L.- COLE, R. (1995) “Spoken Language Corpora”, en COLE, R.A.- MARIANI, J.- USZKOREIT, H.- ZAENEN, A.- ZUE, V. (eds.) (1997) *Survey of the State of the Art in Human Language Technology*, Cambridge: Cambridge University Press, pp. 450-454. Publicación electrónica en URL: <http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html>

LEECH, G. (1991) “The State of the Art in Corpus Linguistics” en AIJMER, K.- ALTENBERG, B. (eds.) *English Corpus Linguistics. Sudies in Honour of Jan Svartvik.* London: Longman, pp. 8-29.

LEECH, G.- BARNETT, R.- KAHREL, P. (1996) *Preliminary Recommendations for the Syntactic Annotation of Corpora*. EAGLES Document EAG-TCWG-SASG1/P-B, March 1996. Publicación electrónica en URL: < http://www.ilc.pi.cnr.it/EAGLES96/segsasg1/segsasg1.html>.

LEECH, G.- FLIGELSTONE, S. (1992) “ Computers and corpus analysis”, en BUTLER, C.S. (ed.) (1992) *Computers and Written Texts*, Oxford: Basil Blackwell, pp. 115-140.

LEECH, G.- MYERS, G.- THOMAS, J. (eds.) (1995) *Spoken English on Computer: Transcription, Markup and Applications*, Harlow: Longman.

LEECH, G.- WILSON, A. (1996) *Recommendations for the Morphosyntactic Annotation of Corpora.* EAGLES Document EAG-TCWG-MAC/R, March 1996. Publicación electrónica en URL: <http://www.ilc.pi.cnr.it/EAGLES96/annotate/annotate.html>.

LEITNER, G. (1992) “International Corpus of English: Corpus Design- problems and suggested solutions”, en LEITNER, G. (ed) *New Directions in English Language Corpora. Methodology, Results, Software Development*, Berlin: Mouton de Gruyter, pp. 33-64.

LEITNER, G. (ed.) (1992) *New Directions in English Language Corpora. Methodology, Results, Software Development*, Berlin: Mouton de Gruyter (Topics in English Linguistics, 9).

LLISTERRI, J. (1994) *Prosody Encoding Survey.* WP 1 Specifications and Standards. T1.5. Markup Specifications. Deliverable 1.5.3. Final version, 15 September 1994. LRE Project 62-050 MULTEXT. Publicación electrónica en URL: <http://www.lpl.univ-aix.fr/projects/multext/CES/CES2.html>.

LLISTERRI, J. (1996a) “Survey of Spanish Resources”, *The ELRA Newsletter*, 1, 1, pp. 7-8. Publicación electrónica en URL: <ftp://ftp.icp.grenet.fr/pub/elra/newslet/en/v1n1/v1n1newsp7.ps.gz>.

LLISTERRI, J. (1996b) *Preliminary Recommendations on Spoken Texts*. EAGLES Documents EAGT C W G - S T P / P , M a y 1 9 9 6 . P u b l i c a c i ó n e l e c t r ó n i c a e n U R L : <http://www.ilc.pi.cnr.it/EAGLES96/spokentx/spokentx.html>.

LLISTERRI, J. (1996c) “Els corpus lingüístics orals”, en CLUB-1, *1er Col.loqui Lingüístic de la Universitat de Barcelona,* Universitat de Barcelona, 20 de desembre de 1993. Secció de Lingüística Catalana, Departament de Filologia Catalana, Universitat de Barcelona.

MacWHINNEY, B. (1991) *The Childes Project: Tools for Analyzing Talk*, Hillsdale, N.J.: Lawrence Erlbaum.

MARINA, J. A.(1993), Teoría de la inteligencia creadora, Barcelona: Anagrama.

McENERY, T.- WILSON, A. (1996) *Corpus Linguistics*, Edinburgh: Edinburgh University Press (Edinburgh Textbooks in Empirical Linguistics).

MEIJS, W. (ed.) (1987) *Corpus Linguistics and Beyond. Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora*, Amsterdam: Rodopi.

MINDT, D. (1996) "English corpus linguistics and the foreign language teaching syllabus", en THOMAS, J.SHORT, M. (Eds) *Using Corpora for Language Research. Studies in Honour of Geoffrey Leech,* London: Longman, pp. 232-247.

OOSTDIJK, N.- de HAAN, P. (eds.) (1994) *Corpus-based Research into Language.In Honour of Jan Aarts*, Amsterdam: Rodopi.

PAYRATÓ, Ll. (1995) "Transcripción del discurso coloquial", en CORTÉS RODRÍGUEZ, L. (ed.) *El español coloquial. Actas del I Simposio sobre Análisis del Discurso Oral.* Almería, 23-25 de noviembre de 1994, Almería: Universidad de Almería, Servicio de Publicaciones, pp. 43-70.

PÉAN, V.- WILLIAMS, S.- ESKÉNAZI, M. (1993) "The Design and Recording of ICY, a Corpus for the Study of Intraspeaker Variability", en *Eurospeech'93. 3rd European Conference on Speech Communication and Technology, Berlin, Germany, 21-23 September 1993*, vol. 1 pp. 627-630.

PINO, M.- SANTALLA, M.P. (1996) "Codificación de la anotación morfosintáctica en SGML", *Procesamiento del Lenguaje Natural, Revista nº* 19, pp. 101-117.

RISSANEN, M.- KYTÖ, M.- PALANDER-COLLIN, M. (eds.) (1993) *Early English in the Computer Age*, Berlin: Mouton de Gruyter.

SÁNCHEZ, A. - SARMIENTO, R.- CANTOS, P.- SIMÓN, J. (1995) *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones,* Madrid: SGEL.

SILVERMAN, K.- BECKMAN, M.- PITRELLI, J.- OSTENDORF, M.- WIGHTMAN, C.- PRICE, P.PIERREHUMBERT, J.- HIRSCHBERG, J. (1992) "TOBI: A standard for labelling English prosody", *Proceedings of the Second International Conference on Spoken Language Processing, ICSLP-92*. Banff, October 1992, pp. 867-870.

SINCLAIR, J. (1996) Preliminary Recommendations on Corpus Typology. EAGLES Document EAGTCWG-CTYP/P, May 1996. Publicación electrónica en URL: <http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html>.

SINCLAIR, J. (Ed) (1987) *Looking Up, An Account of the COBUILD Project.* London: Collins.

SOUTER, C.- ATWELL, E. (eds.) (1993) *Corpus Based Computational Linguistics*. Amsterdam: Rodopi.

SPERBERG-McQUEEN, C.M.- BURNARD, L. (eds.) (1994) *Guidelines for Electronic Text Encoding and Interchange. TEI P3*. Association for Computational Linguistics / Association for Computers and the Humanities / Association for Literary and Linguistic Computing: Chicago and Oxford. Publicación electrónica en URL: <http://etext.virginia.edu/TEI.html>

SVARTVIK, J. (1992) "Corpus linguistics comes of age", en SVARTVIK, J. (ed.) *Directions in Corpus Linguistics (Proceedings of Nobel Symposium 82, Stockholm, 4 - 8, august 1991)*, Berlin - New York: Mouton de Gruyter (Trends in Linguistics), pp. 7 - 13.

SVARTVIK, J. (ed.) (1992) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm 4-8 August 1991*, Berlin: Mouton de Gruyter (Trends in Linguistics, Studies and Monographs, 65)

TAYLOR, L.- LEECH, G. - FLIGELSTONE, S. (1991) A Survey of English machine-readable corpora", en JOHANSSON, S.- STENSTRÖM, A.-B. (eds.) *English Computer Corpora: Selected Papers and Research Guide*, Berlin: Mouton de Gruyter, pp. 319-354.

THOMAS, J.- SHORT, M. (eds.) (1996) *Using Corpora for Language Research. Studies in Honour of Geoffrey Leech,* London: Longman.

TILLMANN, H. G.- POMPINO-MARSCHALL, B. (1993) "Theoretical Principles Concerning Segmentation, Labelling Strategies and Levels of Categorical Annotation for Spoken Language Database Systems" en *Eurospeech'93. 3rd European Conference on Speech Communication and Technology,* Berlin, Germany, 21-23 September 1993, vol. 3, pp. 1691-1694.

VILLENA PONSODA, J.A. (1994) “Pautas y procedimientos de representación del corpus oral de la Universidad de Málaga. Informe preliminar”, en ALVAR EZQUERRA, M.- VILLENA PONSODA, J.A. (Coord) *Estudios para un corpus del español*, Málaga: Universidad de Málaga, pp. 73-102.

WELLS, J.C. (1989) “Computer-coded phonemic notation of individual languages of the European Community “, *Journal of the International Phonetic Association,* 19, 1, pp. 31-54.

WELLS, J.C. (1994) “Computer-coding the IPA: a proposed extension of SAMPA”, *Speech, Hearing and Language, Work in Progress, 1994* (University College London, Department of Phonetics and Linguistics) 8, pp. 271-289.

WINSKI, R. - MOORE, R.- GIBBON, D. (1995) “EAGLES Spoken Language Working Group: Overview and Results”, en *Eurospeech'95. Proceedings of the 4th European Conference on Speech Communication and Speech Technology. Madrid, Spain, 18-21 September, 1995.* Vol 1, pp. 841-844. Publicación electrónica en URL: <http://coral.lili.uni-bielefeld.de/~gibbon/EAGLES/rwpaper/rwpaper.html>.

WINSKI, R.- SENIA, F.- CONNER, P.- HÄB-ÜMBACH, R.- CONSTANTINESCU, A.- NIEDERMAIR, G.- MORENO, A.- TRANCOSO, I. (1996) *Specification of Telephone Speech Data Collection*. LRE-63314 SPEECHDAT, Deliverable D1.4.1. Publicación electrónica en URL: < http://www.icp.grenet.fr/SpeechDat/deliv.html>.