GERARDO ARRARTE
Instituto Cervantes

Normas y estándares para la codificación de textos y para la ingeniería lingüística

El ordenador como herramienta de escritura

Vivimos en una época en que la manipulación de textos en formato digital se ha convertido en una actividad cotidiana para casi todo el inmenso colectivo de personas cuyas actividades se relacionan de forma directa con el lenguaje escrito. Cuantos nos hemos visto alguna vez en esta situación de trabajar con textos mediante la ayuda de un ordenador conocemos las enormes ventajas de las nuevas técnicas sobre los sistemas tradicionales de escritura.

A modo de ejemplo y por recordar sólo algunas de estas ventajas, todos hemos podido comprobar cómo la introducción de cambios en el contenido o en la forma de un documento ya no es una tarea tan laboriosa como antaño, cómo todo texto puede ser fácilmente reproducido o modificado, cómo pueden generarse documentos distintos pero basados en un modelo común de manera automática mediante la combinación de dicho modelo con información contenida en una base de datos, cómo un documento puede ser transmitido o consultado a distancia mediante redes telemáticas, y cómo la búsqueda y recuperación de información se realiza de forma casi instantánea, aun cuando el volumen de la documentación a la que tenemos acceso crece de manera espectacular.

Como cabe esperar, no obstante, no todo son ventajas en el empleo de estas técnicas. Cualquier usuario mínimamente asiduo de las mismas habrá experimentado alguna vez la frustración de ver desaparecer sin remedio un trabajo a causa de un simple error en la manipulación de los documentos archivados en un soporte digital, o bien la desazón de no poder recuperar un texto grabado en un formato que su sistema informático no consigue reconocer.

Es cierto que las técnicas tradicionales de escritura, pese a la labor de los copistas, de los archivos y bibliotecas y de la imprenta, no garantizan tampoco la permanencia de los textos, cuya destrucción accidental o intencionada ha formado parte de la historia humana desde sus orígenes hasta nuestros días. Tampoco resulta siempre una tarea trivial la interpretación de la palabra escrita. Pero el paso de las técnicas mecánicas a las técnicas electrónicas de escritura constituye un auténtico desafío para la conservación de los textos en forma inteligible.

En efecto, en el supuesto de que conozcamos la lengua y el sistema gráfico empleado, podremos descifrar textos codificados mediante técnicas mecánicas, ya se trate de inscripciones sobre piedra o tablas de arcilla, sobre metal, papel

o pergamino, podremos interpretar los jeroglíficos egipcios, los manuscritos del Mar Muerto o las antiguas inscripciones ibéricas, un códice medieval, un incunable del siglo xv o un periódico del día, el texto de una valla publicitaria o el de una pancarta propagandística, el de un anuncio de neón o aquel codificado mediante el sistema Braille. Para ello no necesitamos valernos más que de nuestros propios sentidos y del conocimiento de los códigos empleados en cada caso. Estos códigos, por haber sido diseñados para el uso humano sin ayuda de máquinas, pueden ser dominados tras un aprendizaje adecuado.

¿Podríamos, en cambio, descifrar con igual facilidad los textos codificados por un programa informático y destinados a ser descifrados y manipulados también mediante herramientas informáticas? Estos textos, aun teniendo como destinatario último al hombre, no permiten que éste los lea o consulte de forma directa. En primer lugar, nuestros sentidos son incapaces de percibir sin ayuda las señales que constituyen la representación de un texto en soporte informático.[1] Pero incluso si pudiéramos detectar esas señales, tendríamos ante nosotros unos datos en código binario, es decir una secuencia de dígitos binarios o *bits*.[2] Cada bit representa una cantidad mínima de información: la resultante de la alternancia entre dos estados contrapuestos y complementarios, por lo cual los datos en código binario suelen representarse mediante secuencias de ceros y unos.

Así, en determinados sistemas de codificación digital de textos, cada uno de los caracteres alfabéticos o numéricos que aparecen en esta página, así como los distintos signos de puntuación y espacios entre palabras, estaría representado por una combinación de 8 bits; por ejemplo, en el estándar ISO 8859 la secuencia "01100101" representa la letra "e" minúscula, mientras que "00101101" es la representación del carácter "-". La información codificada de esta manera resulta de difícil manejo para la mente humana y, además, estas largas ristras de ceros y unos que constituyen el texto propiamente dicho suelen ir mezcladas con otros datos, también en código binario, que proporcionan al sistema información complementaria sobre el formato y disposición del texto, sobre el propio sistema de representación empleado o sobre la segmentación de los datos y su distribución en el soporte informático, la cual puede no ser secuencial, sino responder a razones de disponibilidad de espacio.[3]

Debido a esta diferencia cualitativa de los textos codificados mediante sistemas informáticos con respecto a los codificados mediante técnicas mecánicas, aunque los primeros sean fáciles de reproducir y transmitir, nos resulta prácticamente imposible su interpretación si no contamos con la ayuda del

1. Habría que exceptuar algunos soportes ya en desuso, como las tarjetas y cintas perforadas.

2. El término *"bit"* es un acrónimo de la expresión inglesa *"binary digit"*.

3. Los sistemas informáticos suelen almacenar la información de manera similar a la que emplearíamos si, en lugar de escribir un texto de forma continua en un cuaderno, fuéramos pegando notas autoadhesivas de distintos tamaños en las páginas de un álbum, colocándolas allí donde, por haber retirado previamente otras ya inservibles, fueran quedando huecos libres que se adaptaran al tamaño de las nuevas. En este caso, para mantener el orden original de la información, necesitaríamos anotaciones que nos remitieran a la posición de cada retazo de texto en el álbum.

sistema informático adecuado a la codificación en cuestión. La gran variedad de sistemas de codificación textual en soporte informático no sólo dificulta a menudo la interpretación de los textos, sino que puede llegar a amenazar su propia conservación.

Para corroborar esta última afirmación, bastará que el lector con una experiencia suficientemente larga en el uso de sistemas de tratamiento de textos piense en las dificultades que le acarrearía la recuperación de cualquier documento en soporte informático que hubiera creado él mismo hace, por ejemplo, una docena de años. Es muy probable que, entonces, utilizara un ordenador y un programa de procesamiento de textos distintos de los que emplea actualmente. Seguramente, el documento en cuestión estaría almacenado en un disquete flexible de 5,25 pulgadas o en otro soporte cuya lectura requiriese el uso de un dispositivo de uso poco frecuente en la configuración de un sistema informático actual. Aun salvando estas dificultades de lectura física de los datos, es posible que el sistema de codificación resulte total o parcialmente indescifrable para el procesador de textos que use actualmente.

Si consideramos documentos codificados en soporte informático en épocas un poco más lejanas, digamos que anteriores a hace 15 años, las probabilidades de descodificarlo con éxito mediante los recursos técnicos normalmente disponibles en la actualidad disminuyen de manera drástica. Es un hecho que grandes volúmenes de información se han perdido de forma definitiva con los sucesivos cambios en los tipos de soportes y en los sistemas de codificación. Lógicamente, estas pérdidas podrían haberse evitado transfiriendo la información a los nuevos soportes y convirtiéndola a los nuevos códigos; sin embargo, este tipo de precauciones ha sido a menudo la excepción más que la regla.[4]

La normalización de los recursos lingüísticos y de la tecnología lingüística

De lo dicho anteriormente, se deduce fácilmente la necesidad de establecer unas normas para la codificación textual que sean de uso común por parte de distintos usuarios y que, en la medida de lo posible, sean independientes de los sistemas informáticos que se usen. Sin embargo, las necesidades de estandarización en los sistemas de tratamiento del lenguaje abarcan aspectos mucho más variados. La aplicación de sistemas informáticos al estudio y al procesamiento del lenguaje comporta la resolución de problemas tan diversos como la representación en formato digital de caracteres gráficos o señales acústicas, el diseño de corpus de lenguaje escrito o hablado, la elaboración de formalismos léxicos y gramaticales adecuados para conseguir un eficaz tratamiento automático del lenguaje, o la adecuación de determinados módulos de procesamiento lingüístico a su utilización con distintos fines.

Las estrategias adoptadas ante cada uno de estos problemas han sido, a menudo, tan variadas como las características y objetivos de cada proyecto dedicado a la investigación en este campo. Las importantes dificultades que

4. El lector interesado encontrará información más detallada sobre este problema en Rothenberg (1995).

entraña cualquier intento de automatizar los procesos relacionados con algo tan complejo como el lenguaje humano han propiciado la búsqueda constante de soluciones innovadoras.

Ahora bien, el desarrollo de estas soluciones requiere normalmente la utilización de recursos lingüísticos de gran escala. Entendemos por *recursos lingüísticos* todos aquellos materiales en soporte informático, tales como corpus escritos y orales, diccionarios, bases de datos terminológicas, etc., que sirvan como fuente de información, por una parte, para el estudio del lenguaje y, por otra, para el desarrollo de sistemas informáticos que incorporen conocimientos lingüísticos. A estos sistemas que incorporan conocimientos lingüísticos, los llamaremos sistemas de *tecnología lingüística*.

Los distintos tipos de recursos lingüísticos y de sistemas de tecnología lingüística son objeto de análisis detenido en otros capítulos de este libro. En las siguientes páginas se aborda un tema que atañe, en mayor o menor medida, a cada uno de ellos: los esfuerzos llevados a cabo para el desarrollo y la adopción de unas normas aplicables al diseño, a la elaboración y a la utilización de los mismos.

La necesidad de llevar a cabo estos esfuerzos de normalización se debe principalmente al alto coste de estos recursos y sistemas, el cual a menudo impide o, al menos, desaconseja que su diseño se realice teniendo en cuenta únicamente los requisitos planteados por su utilización en el marco de un proyecto concreto o para el desarrollo de un producto específico. En el curso de los últimos años, ha resultado cada vez más evidente la conveniencia de disponer de unos recursos lingüísticos y de una tecnología lingüística genérica que sean compartibles, intercambiables y reutilizables con fines diversos.

Consideraremos que un recurso lingüístico determinado puede ser adecuadamente compartido y reutilizado cuando su diseño se ajuste a unas normas de uso común, o bien permita hacer que se ajuste a ellas estableciendo una correspondencia directa entre su estructura y la requerida por las normas en cuestión. Así, por ejemplo, el hecho de que, en un corpus, tanto la codificación de los textos como la representación de la estructura de los mismos y su anotación lingüística se adapten a determinadas convenciones permitirá que el corpus sea explotado de múltiples formas y mediante cualquier sistema que, a su vez, haya sido diseñado para adaptarse a esas convenciones. Otro tanto ocurrirá con una base de datos léxica o terminológica en cuyas entradas la información adopte una estructura y formato de codificación acorde con unas normas de uso común, o que permitan su adaptación a tales normas.

Del mismo modo, se puede considerar compartible y reutilizable cualquier componente de tecnología lingüística capaz de aceptar unos datos de entrada *(input)* y de producir otros de salida *(output)* que se adapten a unas determinadas normas. Al igual que ocurre con cualquier otro tipo de productos, la adopción de normas y estándares facilita, además, la comparación entre sistemas de características equiparables y la evaluación de la calidad de cada uno de ellos.

Cualquier normalización puede alcanzarse bien mediante el paulatino establecimiento de prácticas comunes y marcos generales compatibles (estándares *de facto*) a través de un consenso que puede ser espontáneo o acordado, o bien

mediante la elaboración de pautas y normas definidas de forma precisa y avaladas por agencias de estándares nacionales o internacionales (estándares *de jure*). Los estándares *de jure* suelen tener su origen en estándares *de facto*, aunque estos últimos no necesariamente acaban por obtener un reconocimiento oficial. Los estándares *de facto* surgen a menudo de una propuesta unilateral de una determinada organización, que obtiene con el tiempo el apoyo de un grupo suficientemente amplio de organizaciones y usuarios. En otros casos, son el resultado de un proceso más formal, que requiere el consenso previo de grupos de empresas, asociaciones profesionales y de usuarios y organismos públicos, los cuales elaboran una propuesta conjunta y la someten a la consideración del resto de personas y entidades que pueden verse afectadas por la definición del estándar, quienes colaboran a su vez en el desarrollo y establecimiento del mismo, tanto con aportaciones que lo modifiquen o amplíen, como con su adopción de forma más o menos generalizada.

En las páginas que siguen, se intenta dar una visión introductoria de tres importantes iniciativas que tienen o han tenido como objetivo la normalización, ya sea en el campo de la codificación de textos o en el de la tecnología lingüística: el estándar *SGML*, la iniciativa *TEI* y el proyecto *EAGLES*. El primero es un *estándar* internacional ya establecido, mientras que los otros dos son proyectos de normalización en marcha actualmente. La TEI ha desarrollado ya unas *normas* que han obtenido una aceptación bastante generalizada entre los investigadores. EAGLES, en cambio, es una iniciativa más reciente, pero que ya ha logrado proponer unas primeras *recomendaciones*.

Tanto la TEI como EAGLES han adoptado como método de trabajo la constitución de grupos de expertos que se han ocupado de distintos aspectos relativos a la codificación textual, en el primer caso, o a la tecnología lingüística, en el segundo, la discusión por parte de estos grupos de las soluciones propuestas a problemas específicos y la elaboración de sucesivas versiones de documentos con recomendaciones, las cuales han sido sometidas a una discusión más amplia por parte de especialistas externos para su posterior refinamiento y actualización.

SGML es un lenguaje que sirve para la representación con carácter general de cualquier tipo de textos en formato digital y su uso está ampliamente extendido en el sector de la edición de documentos electrónicos. La TEI se ocupa de elaborar unas pautas más concretas y específicas para la codificación de tipos textuales determinados, por lo cual sus recomendaciones son de interés para los investigadores y profesionales cuyos trabajos requieran la generación o el empleo de textos en formato electrónico, especialmente cuando estos textos deban trascender el ámbito de trabajo individual y, por tanto, adaptarse a diversas necesidades y al uso de sistemas informáticos variados y cambiantes. Resulta evidente que, entre las disciplinas implicadas en estas labores, la filología y la lingüística ocupan un lugar prominente aunque no exclusivo. El proyecto EAGLES, por su parte, se ocupa de la normalización de los recursos lingüísticos y de los sistemas de tecnología lingüística, por lo cual su ámbito de interés se circunscribe de forma más específica al trabajo de quienes aplican la tecnología informática al estudio del lenguaje y, en especial, de quienes tienen como objetivo el desarrollo de sistemas informáticos de procesamiento del lenguaje.

De lo expuesto se deduce que entre los beneficiarios de las normas y recomendaciones propuestas tanto por la TEI como por EAGLES están, por un lado, quienes diseñan, crean, gestionan y explotan recursos lingüísticos en soporte informático y, por otro, los productores y usuarios de tecnología lingüística.

Comenzaremos por una breve introducción al estándar *ISO 8879*, más conocido como SGML, en el cual se basan las normas TEI, así como buena parte de las recomendaciones de EAGLES.

SGML: Un marco general para la codificación de textos

Cuando escribimos un documento cualquiera con la ayuda de un sistema de procesamiento de textos, introducimos en el ordenador dos tipos de datos: por una parte, cadenas de caracteres que constituyen el texto propiamente dicho y, por otra, *marcas textuales* con las cuales señalamos determinados elementos del documento e indicamos al sistema las funciones de procesamiento que debe llevar a cabo con cada uno de ellos. Podemos definir estas marcas textuales como texto añadido al contenido de un documento con información sobre el mismo.

A veces añadimos estas marcas expresamente; por ejemplo, cuando destacamos un fragmento del texto mediante el uso de una letra de un tipo, estilo o tamaño distinto que el resto, o bien cuando separamos unos párrafos de otros. En otros casos, es el propio sistema el que incorpora las marcas al texto de forma automática; este es el caso cuando se produce un salto automático de línea al llegar al margen derecho. El que la inserción de las marcas textuales se produzca de una u otra forma depende de la capacidad del sistema para tomar la decisión de forma autónoma; así, mientras la separación del texto en párrafos es una decisión propia del autor del documento, el sistema puede estar suficientemente capacitado para tomar decisiones sobre la disposición de las palabras en líneas dentro de un mismo párrafo.[5]

Pues bien, *SGML (Standard Generalized Markup Language)* es un lenguaje de marcas textuales que sirve para representar documentos en formato digital. Por su versatilidad y por tratarse de un estándar internacional oficialmente reconocido y adoptado de forma cada vez más generalizada, SGML cumple de forma satisfactoria el objetivo con que fue diseñado, es decir, que los documentos codificados mediante este lenguaje puedan ser procesados independientemente de los distintos programas, sistemas o dispositivos que se utilicen, de la lengua en que esté escrito el documento, de los juegos de caracteres específicos empleados por distintos sistemas y de la forma de disponer el flujo de datos o la organización física de los archivos.

Para comprender mejor el espíritu que inspiró el surgimiento de este estándar, aprobado en 1986 por la ISO (Organización Internacional para la Estandarización), convendrá repasar brevemente su historia, que comienza por la creación del lenguaje *GML (Generalized Markup Language)* en 1969.

5. A veces, la inserción de una marca textual se produce tras una interacción del sistema con el usuario, como cuando el sistema propone partir una determinada palabra a final de línea y pide la confirmación del autor sobre la adecuación de esa decisión.

A finales de la década de los sesenta, había surgido, entre los especialistas estadounidenses en codificación de textos electrónicos, una tendencia que propugnaba el uso de lo que se denominó *codificación genérica*. Hasta entonces, la codificación de los textos se realizaba mediante la inserción de códigos de control que, al ser detectados por el sistema, hacían que se ejecutaran determinadas instrucciones para que el texto adoptara el formato deseado; estos códigos sólo podían ser interpretados por el sistema para el que habían sido ideados.

En la codificación genérica se sustituían estos códigos de formato por otros de tipo descriptivo; por ejemplo, *"heading"* ("título, encabezamiento") en lugar de *"format-17"*. Aunque pueda parecer trivial, este cambio de unas etiquetas orientadas al funcionamiento del sistema a unas *etiquetas descriptivas* basadas en la estructura del documento conlleva un cambio de enfoque fundamental, que permitirá separar el contenido y la estructura del documento, por una parte, de su formato físico y de los procesos que debe seguir el sistema para producir ese formato, por otra.

El hecho de que las marcas textuales describan la estructura de un documento en lugar de los procesos informáticos que deben llevarse a cabo con él permite que esta codificación descriptiva del documento se realice una sola vez, siendo suficiente para cualquier procesamiento futuro del mismo. Por otra parte, los defensores de la codificación genérica abogaban por el uso de una codificación sistemática y rigurosa que permitiera procesar documentos con las mismas técnicas ya existentes para el procesamiento de otros objetos rigurosamente definidos, tales como programas informáticos o bases de datos.

El lenguaje GML, además de incorporar este enfoque descriptivo, introdujo una novedad importante: la asociación de cada documento a un tipo de documento formalmente definido. Como veremos a continuación, la *declaración o definición*[6] *de tipo de documento (DTD)* característica de SGML consiste básicamente en una descripción explícita de la estructura potencial de un documento perteneciente a un tipo determinado en función de los elementos anidados que pueden formar parte de ella.

El desarrollo de SGML como estándar de codificación textual se produjo a partir de 1978 y culminó con su reconocimiento oficial en 1986. Durante ese período fue adoptado para la codificación de publicaciones y documentos de diversas empresas, instituciones y organismos nacionales e internacionales. Se han desarrollado, asimismo, una variedad de programas informáticos para el tratamiento de documentos SGML, así como diversos conversores que permiten transferir documentos de SGML a otros formatos de codificación y viceversa; algunos de los procesadores de texto más comunes incluyen, en sus versiones más recientes, la posibilidad de convertir los documentos desde y hacia SGML.

6. La bibliografía especializada en SGML y en la TEI habla casi siempre de *"document type definition"*, aunque el texto oficial del estándar ISO 8879 emplea la expresión *"document type declaration"*.

En los últimos años, con la extensión del uso de *hipertextos,*[7] especialmente en los interfaces de usuario de sistemas *multimedia* y para la navegación por Internet, ha alcanzado gran popularidad un lenguaje de codificación hipertextual basado en él: *HTML (HyperText Markup Language).*

Veamos ahora algunas de las características básicas de SGML. En primer lugar, y como puede deducirse de lo dicho anteriormente acerca de las declaraciones de tipos de documentos, el término "documento" tal como se emplea en SGML se refiere no a una entidad física como un archivo o un conjunto de páginas impresas, sino a una entidad lógica que contiene un *elemento documento,* el cual constituye el nodo raíz de un árbol de elementos que forman el contenido del documento. Así, por ejemplo, un documento de un determinado tipo que llamaremos "libro" podrá contener elementos "capítulo", que a su vez podrán contener elementos "sección", y así sucesivamente hasta alcanzar los nodos terminales, que contendrán los caracteres (u otro tipo de datos no textuales)[8] que constituyen el contenido propiamente dicho del documento.

Cada elemento de un documento SGML está delimitado por dos marcas textuales: la *etiqueta inicial* y la *etiqueta final.* Estas etiquetas describen la naturaleza y características del elemento en cuestión, por lo cual reciben el nombre de *marcas descriptivas.* Una de estas características es el tipo de elemento de que se trate, que aparece reflejado en las marcas descriptivas a través de un *identificador genérico.* Además del identificador genérico, las marcas descriptivas pueden contener otros datos, o *atributos,* que aportan información sobre cualidades específicas del elemento descrito.

Las marcas textuales deben ser a su vez debidamente identificadas como tales para distinguirlas del contenido del documento. En el caso de las marcas descriptivas, esto suele[9] hacerse de la siguiente forma: el carácter "<" señala el inicio de una etiqueta inicial y los caracteres "</" el inicio de una etiqueta final; el carácter ">" señala el final de uno u otro tipo de etiqueta.

El ejemplo de la figura 1 incluye varias marcas descriptivas que delimitan los componentes de un fragmento de texto. Los identificadores genéricos empleados en este caso son "tit" (título), "p" (párrafo), "lista", "el" (elemento de una lista) y "td" (texto destacado). En este ejemplo no se han incluido atributos específicos para cada elemento, por lo cual todas las etiquetas iniciales responden al formato "<IG>" y las finales al formato "</IG>", donde "IG" representa al identificador genérico.

7. Se llama *"hipertexto"* a la información estructurada no de forma secuencial, como es habitual en documentos convencionales, sino en forma de red de documentos con vínculos entre sus distintas partes, que los unen de tal forma que se podrían representar mediante un grafo. El lector puede acceder interactivamente a la información que le interesa seleccionando palabras o elementos de un documento que le llevan a otro punto del mismo o de otro documento.

8. Si el nodo terminal es, por ejemplo, un elemento "párrafo", su contenido podrá estar formado por caracteres. En cambio, si se trata de un elemento "figura", es probable que el contenido consista en datos no textuales que constituyan la representación de una imagen.

9. Empleamos aquí la forma de identificar las etiquetas correspondiente a la *sintaxis concreta de referencia.* SGML es, en realidad, un metalenguaje que permite la definición por parte del usuario de distintas *sintaxis concretas.* La sintaxis concreta de referencia, definida en el propio estándar, es la de uso más corriente.


```
[...]
<tit>
Ejemplo de codificaci&oacute;n textual en <td>SGML</td>
</tit>
<p>
He aqu&iacute; un ejemplo de un fragmento de texto codificado en
<td>SGML</td>. Este fragmento consta de un t&iacute;tulo y dos
p&aacute;rrafos, el segundo de los cuales contiene a su vez una
lista de tres elementos. Adem&aacute;s, algunas palabras han sido
destacadas del resto.
<p> Nuestro prop&oacute;sito es:
<lista>
<el>
por una parte, mostrar c&oacute;mo estos componentes del texto
forman parte de su <td>estructura l&oacute;gica</td>, que es
independiente de su <td>estructura f&iacute;sica</td>, la cual
puede ser bien su representaci&oacute;n gr&aacute;fica (mediante
letras y s&iacute;mbolos dispuestos en una p&aacute;gina o pantalla
de acuerdo con diversas convenciones tipogr&aacute;ficas), o bien
la propia codificaci&oacute;n del texto en formato digital
(mediante la disposici&oacute;n lineal de determinadas cadenas de
caracteres definidas tambi&eacute;n de forma convencional);
<el>
por otra parte, ilustrar la correspondencia entre esa &uacute;nica
estructura l&oacute;gica del texto y algunas de sus m&uacute;ltiples
representaciones gr&aacute;ficas posibles;
<el>
finalmente, ofrecer al lector una impresi&oacute;n del aspecto que
puede presentar la codificaci&oacute;n en <td>SGML</td> de un texto
sencillo.
</lista>
[...]
```

Fig. 1. Ejemplo de codificación de un fragmento de texto.

En la figura 1, algunas marcas descriptivas se presentan en líneas separadas. Se trata de una práctica habitual cuyo único fin es facilitar la interpretación humana del texto codificado en SGML, aunque no afecta en nada a la interpretación automática del mismo, ya que el análisis de su estructura se realiza según la disposición de las marcas respecto a la cadena de caracteres que constituyen el contenido del documento y no según su disposición en distintas líneas. También para facilitar la interpretación humana de las marcas, es conveniente emplear nombres que resulten fácilmente inteligibles. Además, aunque aquí hemos empleado en general nombres inspirados en palabras españolas, en algunos casos puede ser preferible usar una nomenclatura más corriente de etiquetas basadas en palabras inglesas que resulte familiar a la mayoría de usuarios de SGML.

Tal como puede observarse en este ejemplo, en algunos casos puede omitirse la etiqueta final; esto es así cuando se ha especificado en la DTD que la misma es opcional. En este caso no se han incluido etiquetas finales para los

elementos de tipo *p* y *el*; se supone, por tanto, que de la información contenida en la DTD se deduce cuándo el final de un elemento dado está implícito por la aparición de otras etiquetas. Por ejemplo, podría darse por concluido un párrafo cuando se inicia otro párrafo, o bien cuando aparece una etiqueta final de un elemento de los tipos sección, capítulo o documento. De igual manera, el final de un elemento de una lista podría estar implícito por el inicio de otro elemento de la lista o por el final de la lista.[10]

La figura 2 muestra algunas de las múltiples representaciones gráficas que podría tomar el fragmento de texto cuya codificación hemos visto. Puesto que la codificación es puramente descriptiva, el aspecto y disposición del texto depen-derán de las correspondencias que se establezcan entre la estructura del texto y las funciones que deba ejecutar el sistema para cada elemento de esa estructura.

En estos ejemplos no se muestra la estructura general del documento; aunque se aprecia la forma en que están anidados los elementos dentro de la lista, ésta dentro de un párrafo y los fragmentos de texto destacado dentro de algunos de estos elementos, no sabemos de qué manera dependen los elementos de tipo *tit* y *p* de los elementos que constituyen los nodos superiores del árbol hasta llegar al nodo raíz.

Veamos, por tanto, en la figura 3 un ejemplo que nos permita tener una visión global de la estructura de un documento completo como puede ser un libro (aunque en este caso se trata de una estructura más bien simple en relación con la de un libro típico). Vemos que la totalidad del documento consta de un elemento que hemos llamado *libro*, cuyos componentes inmediatos son tres elementos denominados *prelim, cuerpo* y *post*. El elemento *prelim* contiene a su vez varios elementos preliminares como el título, el autor y el prólogo, mientras que el elemento *cuerpo* consta de una serie de capítulos, y el elemento *post*, de una serie de apéndices. Tanto el prólogo, como cada uno de los capítulos y de los apéndices están compuestos por párrafos; los capítulos y los apéndices, además, comienzan con un título. Los nodos terminales de la estructura del documento son, por lo tanto, los denominados *titulo*,[11] *autor*, *p* (párrafo) y *tit* (título de un capítulo o apéndice). El contenido de estos nodos terminales, indicado en el ejemplo mediante puntos suspensivos, sería el contenido del documento y podría estar constituido en este caso por simples cadenas de caracteres.

Todos los ejemplos de marcas descriptivas vistos hasta el momento contienen solamente un identificador genérico, el cual, como se ha dicho, indica de qué tipo de elemento se trata. Hemos visto, sin embargo, que las marcas descriptivas pueden contener también *atributos*, es decir, datos que nos informan sobre cualidades específicas del elemento concreto identificado por la marca.

10. Conviene tener presente que se trata sólo de un ejemplo; estas suposiciones no tienen valor general, sino que dependerán siempre de la DTD asociada a cada documento.

11. Debemos señalar que la ausencia de tildes en las etiquetas de los ejemplos no significa que SGML no permita hacer uso de caracteres acentuados en las marcas textuales. No obstante, es práctica común ceñirse a un juego de caracteres restringido con el fin de no complicar innecesariamente la sintaxis empleada en la codificación. Hay que tener en cuenta que esto no afecta al contenido del documento.

[...]

Ejemplo de codificación textual en *SGML*

He aquí un ejemplo de un fragmento de texto codificado en *SGML*. Este fragmento consta de un título y dos párrafos, el segundo de los cuales contiene a su vez una lista de tres elementos. Además, algunas palabras han sido destacadas del texto.

Nuestro propósito es:

- por una parte, mostrar cómo estos componentes del texto forman parte de su *estructura lógica*, que es independiente de su *estructura física*, la cual puede ser bien su representación gráfica (mediante letras y símbolos dispuestos en una página o pantalla de acuerdo con diversas convenciones tipográficas), o bien la propia codificación del texto en formato digital (mediante la disposición lineal de determinadas cadenas de caracteres definidas también de forma convencional);
- por otra parte, ilustrar la correspondencia entre esa única estructura lógica del texto y algunas de sus múltiples representaciones gráficas posibles,
- finalmente, ofrecer al lector una impresión del aspecto que puede presentar la codificación en *SGML* de un texto sencillo.

[...]

[...]

Ejemplo de codificación textual en SGML

He aquí un ejemplo de un fragmento de texto codificado en SGML. *Este fragmento consta de un título y dos párrafos, el segundo de los cuales contiene a su vez una lista de tres elementos. Además, algunas palabras han sido destacadas del resto.*

Nuestro propósito es:

i) por una parte, mostrar cómo estos componentes del texto forman parte de su estructura lógica, *que es independiente de su* estructura física, *la cual puede ser bien su representación gráfica (mediante letras y símbolos dispuestos en una página o pantalla de acuerdo con diversas convenciones tipográficas), o bien la propia codificación del texto en formato digital (mediante la disposición lineal de determinadas cadenas de caracteres definidas también de forma convencional);*

ii) por otra parte, ilustrar la correspondencia entre esa única estructura lógica del texto y algunas de sus múltiples representaciones gráficas posibles;

iii) finalmente, ofrecer al lector una impresión del aspecto que puede presentar la codificación en SGML *de un texto sencillo.*

[...]

Fig. 2. Ejemplos de distintas representaciones gráficas de un fragmento de texto.

Sólo a modo de ejemplo, mencionaremos uno de los atributos más habituales: el que sirve para identificar de forma única a un determinado elemento del documento. El documento cuya estructura acabamos de analizar (figura 3) contiene algunos tipos de elementos que sólo aparecen una vez (*libro, prelim, titulo,*

autor, prologo, cuerpo y *post*). En cambio, los tipos de elementos *cap, apend, tit* y *p* pueden aparecer un número indefinido de veces. Puede resultar de especial interés que las marcas descriptivas de un tipo de elemento puedan incluir un identificador único de un elemento concreto de ese tipo. Esto permitiría, por ejemplo, realizar referencias al contenido de ese elemento en otros lugares del documento.

```
<libro>
   <prelim>
         <titulo>...</titulo>
         <autor>...</autor>
         <prologo>
               <p>...
               <p>...
               .
               .
               .
               <p>...
         </prologo>
   </prelim>
   <cuerpo>                                  <post>
         <cap>                                     <apend>
               <tit>...</tit>                            <tit>...</tit>
               <p>...                                    <p>...
               <p>...                                    <p>...
               .                                         .
               .                                         .
               .                                         .
               <p>...                                    <p>...
         <cap>                                     <apend>
               <tit>...</tit>                            <tit>...</tit>
               <p>...                                    <p>...
               <p>...                                    <p>...
               .                                         .
               .                                         .
               .                                         .
               <p>...                                    <p>...
            .                                         .
            .                                         .
            .                                         .
         <cap>                                     <apend>
               <tit>...</tit>                            <tit>...</tit>
               <p>...                                    <p>...
               <p>...                                    <p>...
               .                                         .
               .                                         .
               .                                         .
               <p>...                                    <p>...
</cuerpo>                                 <post>
                                       </libro>
```

Fig. 3. Ejemplo de codificación en SGML de un libro de estructura muy simple.

Así, podemos añadir a las marcas descriptivas de elementos como *cap* y *p* un atributo que llamaremos *"id"* y cuyo valor será cualquier cadena de caracteres que no se repita como valor del atributo *id* de ningún otro elemento del documento y que, además, se ajuste a otras restricciones que especifiquemos para el valor de dicho atributo en lo que concierne a la longitud de la cadena y al conjunto de caracteres que pueden formar parte de ella.

En el ejemplo de la figura 4 hemos añadido este atributo en las etiquetas iniciales de algunos de los elementos de tipo *cap* y *p*. Esto nos permitirá realizar referencias a los elementos así identificados.

```
[...]
<cap id=estrvege>
   <tit>Estructura y medio ambiente de la vegetaci&oacute;n</tit>
   <p>
   La vegetaci&oacute;n que cubre los continentes de la Tierra
   es de vital importancia [...]
   <p id=dfvegnat>
   El tema que aqu&iacute; nos ocupa es el de la <td>vegetaci&oacute;n
   natural</td>, es decir, la vegetaci&oacute;n que se desarrolla sin
   ser interferida y modificada apreciablemente por el hombre. [...]
   .
   .
   .
<cap id=distrvege>
   <tit>Distribuci&oacute;n de la vegetaci&oacute;n natural</tit>
   <p>
   En el cap&iacute;tulo <capref refid=estrvege> ya se han tratado
   los principios de descripci&oacute;n de la vegetaci&oacute;n en
   funci&oacute;n de su estructura, [...]
   <p>
   Toda la vegetaci&oacute;n natural (v&eacute;ase <pref refid=dfvegnat>)
   puede agruparse en cuatro grandes subdivisiones estructurales,
   [...]
   p id=dfbosque>
   Un <td>bosque</td> puede definirse como una formaci&oacute;n vege-
   tal constituida por &aacute;rboles que crecen unos junto a otros y
   forman un estrato de hojas que cubre de sombra el suelo.
[...]
```

Fig. 4. Ejemplo de codificación con uso de atributos de identificación única de elementos

Al editar un documento mediante un lenguaje de codificación genérica como SGML, es imposible conocer la disposición final del texto en páginas, ya que ésta será el resultado del proceso que se lleve a cabo con el texto. Por ello, la posibilidad de referirnos a cualquier elemento independientemente de su ubicación final en el documento resulta especialmente útil. Incluso el orden y la numeración de los capítulos y secciones, así como el de las ilustraciones, pueden sufrir numerosos cambios a lo largo de la edición del documento. Es

por esto que resulta conveniente el uso de identificadores que, como los del ejemplo de la figura 4, no dependan de esa numeración. Aunque también es práctica habitual el uso de identificadores basados en la numeración (por ejemplo, "<cap id=cap20>"), es aconsejable el empleo de nombres que nos remitan al contenido del elemento en cuestión más que a su situación en el documento.

Los sistemas que se utilicen para procesar el documento posteriormente a su codificación en SGML se encargarán de distribuir el texto en páginas y de numerar tanto las páginas como otros elementos del documento; asimismo, podrán asignar a cada referencia el formato que se estime oportuno. En la figura 5 se muestran algunas de las diversas formas que podrían tomar los fragmentos del texto de la figura 4 que incluyen referencias a otros elementos del documento.

Este ejemplo nos ha llevado a la introducción de un tipo excepcional de marcas descriptivas: las empleadas para indicar un *elemento vacío*, es decir, un elemento cuyo contenido no es introducido por el usuario, sino que es generado por el sistema de procesamiento. Los elementos *capref* y *pref* usados en el ejemplo para introducir referencias a capítulos o párrafos son elementos de este tipo; presentan una estructura similar a las etiquetas iniciales de un elemento: un indicador genérico ("capref" y "pref"), un atributo ("refid") con la indicación de su valor y los delimitadores inicial ("<") y final (">"). Como cabe esperar, las marcas de este tipo nunca van acompañadas de una etiqueta final.

```
[...]
En el cap&iacute;tulo <capref refid=estrvege> ya se han tratado [...]

    En el capítulo 20 ya se han tratado...
    En el capítulo 20, Estructura y medio ambiente de la vegetación, ya
    se han tratado...
    En el capítulo Estructura y medio ambiente de la vegetación (página
    351) ya se han tratado...
    En el capítulo Estructura y medio ambiente de la vegetación
    pp. 351-367) ya se han tratado...

Toda la vegetaci&oacute;n natural (v&eacute;ase <pref refid=dfvegnat>)
puede agruparse [...]

    Toda la vegetación natural (véase pág. 351) puede agruparse...
    Toda la vegetación natural (véase pág. 351, párrafo 2) puede
    agruparse...
    Toda la vegetación natural (véase Capítulo 20, pág. 351) puede
    agruparse...

[...]
```

Fig. 5. Algunas de las formas que pueden tomar las referencias del ejemplo anterior (figura 4).

Además de las marcas descriptivas que, como hemos visto, se emplean para delimitar los elementos que componen la estructura de un documento, se emplean en SGML otros tipos de marcas. Uno de ellos es el constituido por las *referencias a entidades*. En un sistema informático determinado, un documento puede estar almacenado en varias partes, cada una de ellas en una unidad de almacenamiento separada (archivos, conjuntos de datos, miembros de "librerías" o bibliotecas de datos, etc.). A cada una de esas partes se las llama *entidades*. Estas entidades están conectadas por las referencias a entidades que forman parte de la codificación del documento. Una referencia a una entidad es una llamada para que el sistema inserte un determinado texto (la entidad) en el documento en el lugar de aparición de la referencia. El texto que constituye la entidad puede haber sido definido dentro del mismo documento o bien de forma externa.

El lector habrá observado ya en los ejemplos precedentes el uso de determinadas secuencias de caracteres en sustitución de algunos caracteres especiales, como las letras acentuadas y la letra "ñ". Secuencias como "é", "&uumlaut;" o "ñ" son referencias a las entidades "é", "ü" y "ñ". Los caracteres "&" y ";" actúan como delimitadores de las referencias a entidades, permitiendo distinguirlas del resto de caracteres del documento. Para que estas referencias surtan el efecto deseado durante el procesamiento del documento, es necesario que las entidades estén definidas en algún sitio. En los ejemplos anteriores, las entidades pertenecen a uno de diversos *conjuntos públicos de entidades*, es decir a un conjunto de entidades empleado de forma estándar o compartido por una determinada comunidad de usuarios para la codificación de documentos. SGML nos permite definir nuestras propias entidades, por lo cual podríamos haber optado por nombres como "&eacento;", "&udieresis;" o "&enne;". No obstante, el uso de entidades pertenecientes a conjuntos públicos facilita la interpretación del documento por parte de otros usuarios y sistemas sin necesidad de recurrir a un nuevo juego de entidades definido especialmente para ese documento. También podríamos haber prescindido por completo del uso de estas referencias a entidades empleando, previa declaración del mismo en la DTD del documento, un conjunto de caracteres que incluyera los de uso más frecuente en el texto, pero esto podría causar igualmente problemas a usuarios y sistemas que no utilizaran ese mismo juego de caracteres.

Las referencias a entidades no sólo son útiles para sustituir caracteres que, como éstos, pueden no estar disponibles en determinado teclado, no ser visualizables en pantalla en determinado sistema o presentar problemas en su transmisión de un sistema informático a otro. También permiten emplear abreviaturas en sustitución de largas cadenas de caracteres. Por ejemplo, si en un documento se repite a menudo la expresión "Organización de las Naciones Unidas para la Alimentación y la Agricultura", podríamos declararla como una entidad y sustituirla en el texto por la referencia "&fao;". Finalmente, mediante referencias a entidades se puede indicar el lugar de inserción de una parte del documento almacenada en un archivo distinto.

Aparte de las marcas descriptivas y de las referencias a entidades, existen otros dos tipos de marcas en SGML: las *declaraciones*, que definen el uso de las restantes marcas y controlan su interpretación, que se usan fundamentalmente

en las declaraciones de tipos de documentos (DTD); y, finalmente, las *instrucciones de procesamiento*, que, a diferencia de las restantes marcas, se codifican en función de un sistema informático concreto y constituyen, por lo tanto, un último recurso cuando el uso de codificación genérica no resulta adecuado para nuestros fines.

Para concluir esta breve introducción a SGML, veremos de forma somera en qué consiste la *declaración de tipo de documento (DTD)*. Una DTD contiene, entre otras informaciones, las reglas que definen las estructuras permitidas para cualquier documento de un tipo dado. Los elementos que componen un documento pueden aparecer en él sólo de acuerdo con esas reglas, que se definen en las *declaraciones de elementos*.

La figura 6 muestra las declaraciones de algunos de los elementos que hemos visto en nuestro ejemplo de estructura de un libro sencillo (figura 3). El delimitador inicial "<!" indica que se trata de declaraciones, y la palabra clave "ELEMENT" que aparece a continuación, el tipo de declaración. Los dobles guiones que acompañan a los delimitadores de la primera línea indican que el contenido de esta marca constituye un *comentario*, es decir, un texto explicativo para referencia del usuario pero que será ignorado por el sistema. Para facilitar la comprensión de los códigos, éstos se han dispuesto en forma de tabla mediante tabuladores; el sistema ignorará también esta disposición: en las marcas de SGML, varios espacios (o códigos de tabulación) consecutivos equivalen a un solo espacio. El nombre que sigue a "ELEMENT", y que aquí hemos colocado bajo el epígrafe "Elementos", es el identificador genérico del elemento que se declara, mientras que los nombres alineados bajo el epígrafe "Contenido" indican los identificadores genéricos de los elementos que puede contener el elemento declarado.

```
<!--        ELEMENTOS          MIN     CONTENIDO                    -->
<!ELEMENT   libro              - -     (prelim?, cuerpo, post?)     >
<!ELEMENT   cuerpo             - o     cap+                         >
<!ELEMENT   (cap | apend)      - o     (tit, (p | fig)+)            >
```

Fig. 6. Ejemplos de declaraciones de elementos

Así, la primera declaración corresponde al elemento *libro* y establece que su contenido constará de un elemento *prelim*, seguido por un elemento *cuerpo*, seguido a su vez por un elemento *post*. Las comas (",") que separan los tres identificadores genéricos indican que se trata de una secuencia ordenada, es decir, que no pueden aparecer en otro orden; el signo "?" a continuación de los elementos *prelim* y *post* indica que ambos son optativos. El signo "+" en la declaración del elemento *cuerpo* indica que éste debe estar formado por *uno o más* elementos *cap*. La barra vertical ("|") de la tercera declaración es un operador disyuntivo: su primera aparición indica que la declaración es válida para el elemento *cap* o para el elemento *apend*; en el grupo de nombres que define

el contenido permitido para estos dos elementos, se ha usado la barra vertical para indicar que al elemento *tit* deben seguir uno o más elementos pertenecientes al par formado por *p* y *fig*.

Los dos caracteres situados en el ejemplo bajo el epígrafe "Min" y comprendidos entre el nombre o grupo de nombres de los elementos declarados y el nombre o grupo de nombres de los elementos del contenido corresponden respectivamente a los dos *parámetros de minimización* de las etiquetas inicial y final y sirven para indicar en qué casos se puede omitir alguna de estas etiquetas. La letra "o" indica que la omisión está permitida, mientras que el signo "-" indica que no lo está. Por lo tanto, en nuestro ejemplo, podremos omitir la etiqueta final de los elementos *cuerpo*, *cap* y *apend*. La omisión de la etiqueta final del elemento *cuerpo* es posible debido a que el final de este elemento siempre coincide con el inicio de *post* o, en caso de que no aparezca este elemento optativo, con el final del elemento *libro*. Para los elementos *cap* y *apend*, la posibilidad de omitir la etiqueta final se debe a que el final de cualquier elemento de uno de estos tipos está implícito bien por el inicio de otro elemento del mismo tipo, o bien por el final del elemento que lo contiene (*cuerpo* y *post*, respectivamente). La omisión no sería posible en elementos que pudieran contener elementos del mismo tipo, permitiendo por tanto anidar elementos con un mismo identificador genérico (por ejemplo, una cita dentro de otra cita, o una lista de elementos dentro de otra).

Además de las declaraciones de elementos, la DTD debe contener otra información, como las declaraciones de las listas de atributos asociados a cada elemento y las declaraciones de las entidades empleadas (ya sea de aquellas definidas por el usuario o de las pertenecientes a conjuntos públicos de entidades, con indicación, en este último caso del conjunto de que se trata). Los ejemplos anteriores constituyen sólo una primera aproximación a la codificación de documentos en SGML, por lo cual remitimos al lector interesado a la bibliografía indicada.[12]

Conviene realizar una aclaración final en relación con las diversas aplicaciones de SGML y con el enfoque adoptado en esta sección. Los ejemplos empleados y la mayor parte de las explicaciones han servido para describir este lenguaje de codificación desde la perspectiva de quien crea un documento y lo codifica para su posterior procesamiento mediante sistemas informáticos. Es necesario tener en cuenta que SGML se emplea también, y muy especialmente en el campo de la filología y la lingüística, para la codificación de documentos ya existentes, tanto en la preparación de ediciones críticas de los mismos como en su disposición en soporte informático para la realización de estudios estilísticos o para su uso como parte de corpus lingüísticos. En estos casos el codificador, en lugar de generar una estructura de marcas que se conforme a su concepción del documento que está creando, generará una que se adecue al documento original y que refleje lo más fielmente posible su interpretación del mismo.

12. Goldfarb (1990) contiene el texto oficial completo del estándar ISO 8879, acompañado de abundantes aclaraciones, ejemplos, índices y referencias cruzadas que facilitan su comprensión.

TEI: Unas normas comunes para la codificación y el intercambio de textos

Es precisamente esta necesidad, indicada en el último párrafo de la sección anterior, de codificar textos ya existentes, o bien de permitir el intercambio de textos ya codificados electrónicamente, uno de los motivos fundamentales que dieron origen a la iniciativa *TEI (Text Encoding Initiative)*. Como hemos visto, el lenguaje SGML proporciona al codificador de textos en formato electrónico un marco general que le deja suficiente autonomía para crear su propio esquema de codificación según sus necesidades y las características del texto en cuestión. La TEI va más allá: su objetivo es la elaboración de unas normas que aborden en su integridad el vasto problema planteado por la representación de información textual en formato electrónico y que tengan en cuenta la multiplicidad de fenómenos que pueden presentarse en distintos tipos de texto, así como la diversidad de fines con que la representación de dichos fenómenos puede ser usada, especialmente en los distintos campos de investigación que se basan en el análisis de textos.

La historia de esta iniciativa comienza en noviembre de 1987, cuando un grupo de unos 30 expertos provenientes de archivos, centros de investigación y asociaciones profesionales se reúne en Vassar College (Poughkeepsie, Nueva York) para tratar el problema de la estandarización en un área que, durante las últimas décadas, había visto surgir una multitud de esquemas distintos y a menudo incompatibles que intentaban hacer frente a la necesidad de representar caracteres especiales, codificar las divisiones lógicas de un texto, representar información analítica o interpretativa, o reducir todo el conjunto de informaciones de crítica textual con las distintas interpretaciones y anotaciones de un texto a una única secuencia lineal.

Esta iniciativa de 1987 había sido precedida ya por otros intentos de estandarización, aunque fue en esta ocasión cuando por primera vez se alcanzó un consenso sobre unos principios que servirían de guía al proceso que se llevaría a cabo en los años siguientes. La urgencia del problema, cuyas consecuencias se hacían sentir con más apremio año a año, el compromiso de las principales organizaciones y centros de investigación especializados en la manipulación de textos en formato electrónico, así como las posibilidades abiertas por la recien-te aprobación del estándar SGML, fueron seguramente causa del éxito del encuentro.

De este encuentro nace, pues, la TEI como iniciativa conjunta de tres asociaciones profesionales, la *ACH (Association for Computers and the Humanities)*, la *ALLC (Association for Literary and Linguistic Computing)* y la *ACL (Association for Computational Linguistics)*, a las cuales se sumarían diversos organismos de investigación y proyectos afiliados así como otras entidades públicas y privadas de Europa y Norteamérica. Si bien la TEI surgía como un proyecto internacional y plurilingüe cuyo objetivo era el desarrollo de normas comunes para la codificación y el intercambio de textos electrónicos con fines de investigación científica, pronto resultaría evidente que sus objetivos eran de interés primordial para aplicaciones de todo tipo en el creciente mundo de las industrias de la lengua.

El método de trabajo adoptado por la TEI se ha basado en el estudio minucioso, por parte de grupos de especialistas en cada área, de las distintas cuestiones concernientes a la codificación de textos de tipos muy diversos, en la elaboración de propuestas pormenorizadas pero flexibles para el tratamiento de esas cuestiones y en una amplia discusión de las sucesivas versiones de las recomendaciones resultantes por parte de la comunidad científica en general.

Las normas de la TEI[13] proporcionan convenciones de codificación que permiten describir la estructura física y lógica de una gran variedad de tipos de textos, así como elementos característicos de tipos de textos determinados, abordando los problemas habituales que surgen durante la codificación textual.

La TEI ofrece soluciones para un amplio espectro de usuarios, que incluye a investigadores de campos como las humanidades, las ciencias sociales y otras muchas disciplinas científicas, a editores, bibliotecarios y documentalistas. También da respuestas a muchas de las necesidades del creciente sector de la tecnología lingüística, en el cual se están constituyendo importantes corpus de textos escritos y de lengua oral, así como diccionarios computacionales y bases de datos terminológicas, con el fin de avanzar en el desarrollo de sistemas capaces de manipular el lenguaje humano, trabajos que, como hemos visto, requieren cada vez más unas normas comunes que permitan que sean compartibles e intercambiables.

Los grupos de trabajo encargados del estudio de los distintos aspectos específicos de la codificación textual dependen de cuatro comités que se ocupan de las grandes áreas abordadas en este proyecto: la documentación de los textos, su representación, su análisis e interpretación y las cuestiones metalingüísticas. A continuación, mencionaremos muy brevemente el trabajo realizado en cada una de estas áreas.

El comité encargado de la *documentación de los textos* trabaja en la elaboración de las recomendaciones relativas al *prólogo o cabecera*, es decir, a aquella sección de cualquier documento codificado conforme a la TEI que contiene información sobre el mismo. Esta información metatextual incluye, por ejemplo, la identificación y descripción bibliográfica del texto codificado que permita al usuario localizar la edición empleada en su codificación o bien alguna otra edición del mismo. La cabecera debe contener, asimismo, la identificación de la propia codificación del texto que permita a bibliotecarios y documentalistas catalogar los archivos informáticos que la contienen, así como otra información de interés para los gestores de los archivos o para sus usuarios. Finalmente, la cabecera de un documento TEI contiene las declaraciones relativas al propio sistema de codificación empleado, para que los programas puedan interpretar las marcas textuales usadas y procesar el texto de forma adecuada.

El comité de *representación textual* se encarga de proporcionar recomendaciones para la adecuada representación de las versiones impresas o manuscritas

13. El texto completo de estas normas en su estado actual se recoge en Sperberg-McQueen & Burnard (1994). Este documento, que comprende dos volúmenes y más de 1.300 páginas, es conocido comúnmente por el nombre abreviado de *TEI P3 (TEI Proposal 3)* por tratarse de la tercera versión de estas recomendaciones.

del texto. Esta representación incluye tanto la descripción física del texto según su disposición en la edición de la que fue tomado para su codificación, como la descripción lógica de los elementos del texto representados mediante convenciones tipográficas, tales como los caracteres especiales, símbolos y caracteres correspondientes a otros alfabetos, los elementos de la jerarquía estructural del texto (*v.g.* libro, capítulo, verso), otros elementos habituales en un texto representados mediante recursos tipográficos (*v.g.* texto destacado, citas, disposición tabular) o elementos menos habituales que pueden acompañar a un texto (*v.g.* notas, anotaciones marginales, aclaraciones, textos paralelos, rectificaciones editoriales, comentarios de crítica textual).

El comité de *análisis e interpretación* tiene como cometido el estudio de la codificación de aquellos elementos implícitos en el texto que no suelen representarse tipográficamente. Algunos de los problemas abordados son comunes a diversos campos, tales como las referencias intratextuales e intertextuales, la delimitación de determinados segmentos de texto con referencias a comentarios y a otros materiales relacionados, o las etiquetas que permiten la inclusión de elementos o segmentos del texto en índices y bajo términos determinados; otros, en cambio, corresponden al campo específico del análisis lingüístico del texto, como los relativos a su análisis sintáctico, morfológico o léxico, a la anotación de corpus lingüísticos o a la codificación de diccionarios; un tercer grupo corresponde a los estudios literarios e incluye las etiquetas para estudios temáticos, para la identificación de alusiones, así como marcas textuales ideadas especialmente para la representación de elementos propios de determinados tipos textuales y géneros literarios.

Como cabe suponer, la representación de aspectos analíticos e interpretativos del texto plantea más problemas que la de aquellos elementos que ya están señalados en él mediante convenciones tipográficas. La necesidad de definir conjuntos de etiquetas para diversos campos requiere tomar las máximas precauciones para no caer en presuposiciones teóricas que harían que el esquema de codificación no resultase aceptable para los investigadores que no compartieran dichas presuposiciones, por lo cual se ha intentado definirlos de tal manera que permitieran la codificación de los fenómenos considerados de interés por distintas teorías. Para ello, se han delimitado las áreas que se prestan a divergencias en el análisis del texto, permitiendo en algunos casos al codificador la declaración del uso de prácticas específicas en esas áreas o bien, en otros casos, unificando las distintas posiciones existentes en un único conjunto de etiquetas neutral y válido para diferentes teorías.

Finalmente, el cuarto comité, el responsable de *cuestiones metalingüísticas*, está a cargo de elaborar una sintaxis concreta adecuada para los conjuntos de etiquetas propuestos por la TEI. Desde un principio, se determinó que el lenguaje SGML constituía el marco apropiado para el desarrollo de la TEI. Se establecieron algunas restricciones y pautas sobre su uso para hacer frente con éxito al intercambio entre sistemas diferentes, si bien se procuró mantener el carácter general y flexible de SGML que le permite adecuarse a una amplia gama de necesidades.

El esquema de codificación de la TEI ha sido diseñado teniendo en cuenta los siguientes objetivos: que resulte suficiente para representar los elementos del

texto necesarios para los investigadores, sin perder por ello de vista las prácticas habituales y las necesidades de las editoriales y de los productores de programas comerciales; que sea simple, claro y concreto; que sea de fácil utilización y que no requiera el uso de programas específicos; que permita una definición rigurosa y un procesamiento eficiente de los textos; que permita al usuario la definición de extensiones propias; y que se ajuste a estándares existentes o en desarrollo.

La búsqueda de claridad y concreción ha llevado a que, lejos de limitarse a dar recomendaciones generales sobre la construcción de una DTD para la codificación de documentos conforme a las normas TEI o a ofrecer un modelo abstracto de DTD, el esquema de codificación incluya la especificación de una DTD completa. Esto permite el uso de dicho esquema por parte de investigadores sin conocimientos previos sobre esta cuestión, sin por ello impedir la modificación y ampliación del mismo cuando se desee adaptarlo a fines específicos.

El afán de simplicidad ha conducido a evitar, en la medida de lo posible, la proliferación de elementos, haciendo coincidir en uno solo aquellos elementos que presentan ciertas semejanzas. Así, por ejemplo, mientras algunos lenguajes de codificación emplean marcas distintas para tres o más tipos de listas,[14] la TEI define un único elemento "`<list>`" y deja las distinciones tipológicas a un atributo especial, "`type`". Sin embargo, este intento de evitar la proliferación de elementos ha debido dejar paso a menudo a una tendencia contraria debida a la necesidad de dar soluciones simples a los casos simples. En efecto, además de una notación general, conviene a veces poder disponer de marcas más específicas para manejar esos casos simples mediante una notación más sencilla. Este tipo de marcas, si bien resultan redundantes y contradicen, por tanto, el principio general que acabamos de enunciar, permiten adaptar el grado de complejidad del esquema de codificación al grado de complejidad con que deseemos describir un texto.

Puesto que el esquema de codificación debe poder extenderse a casos imprevistos, ya sea por las características peculiares del texto o por las del tipo de anotación que deseemos realizar, la TEI prevé la posibilidad de extender y modificar el modelo de DTD propuesto, el cual tiene, como veremos, una estructura modular que consta de un conjunto de marcas básico y distintos conjuntos adicionales de marcas especializadas. Además, el codificador puede introducir módulos nuevos, suprimir determinadas declaraciones, cambiar el nombre de elementos existentes o bien añadir otros nuevos.

Los requisitos establecidos para que la codificación de un texto se conforme a las normas TEI varían según se trate de una codificación nueva o bien de la traducción de textos ya existentes en formato electrónico al esquema de codificación propuesto por la TEI. En este último caso, y para facilitar esa conversión al formato común de la TEI, se considera suficiente que no se

14. Un determinado lenguaje de codificación emplea, por ejemplo, las marcas "`:ol.`" y "`:eol.`" (del inglés *"ordered list"* y *"end of ordered list"*, respectivamente) para aquellas en que los elementos son introducidos por un símbolo numérico o alfabético que indica su orden, "`:ul.`" y "`:eul.`" (inglés, *"unordered list"*) para aquellas en que todos los elementos están marcados por un mismo símbolo, como un punto elevado u otro carácter semejante, y "`:sl.`" y "`:esl.`" (*"simple list"*) para las que carecen de unas u otras señales.

produzca pérdida alguna de información. En cambio, para las codificaciones nuevas, se establecen unos mínimos en relación con la información que debe ser representada.

Veamos brevemente algunas de las principales características del sistema de marcas textuales propuesto por la TEI. En primer lugar, y como ya se ha dicho al tratar la codificación genérica, hay que distinguir entre la estructura física y la estructura lógica del texto: un sistema de marcas que describe meramente la presentación del texto puede usar una etiqueta para señalar el hecho de que un fragmento esté impreso en cursiva, mientras que un sistema de codificación genérica tiende a identificar fenómenos algo más abstractos, como el hecho de que el fragmento en cuestión sea, por ejemplo, una cita metalingüística de una determinada palabra o expresión o bien un fragmento en una lengua distinta a la empleada en el texto. La TEI, a pesar de adoptar un enfoque genérico, permite la representación de la estructura física del texto, ya que ésta puede ser fundamental para la investigación en determinados campos.

Una *etiqueta* es una marca textual específica que señala la presencia o la localización de un determinado fenómeno en el texto. Un mismo fenómeno puede ser indicado por etiquetas distintas en distintos esquemas de codificación. Aunque la TEI propone un modelo de DTD basado en conjuntos de etiquetas predefinidos, el codificador dispone, si lo desea, de mecanismos para abreviar o traducir a una determinada lengua los identificadores genéricos de las etiquetas sin alterar la definición de los fenómenos que señalan o la sintaxis que rige su aparición en el texto.

Las normas TEI describen varias declaraciones de tipos de documentos (DTD): una única DTD principal para la transcripción de textos y varias auxiliares para la codificación de la información metatextual relativa a la transcripción de uno o más textos. Una DTD auxiliar comprende: la cabecera independiente, en la cual se documenta la identidad de un determinado texto electrónico y de su fuente; la declaración de sistema de escritura, en la cual se describe el alfabeto o sistema de escritura, así como los conjuntos de caracteres, esquemas de transliteración y conjuntos de entidades SGML empleados; la declaración del sistema de fenómenos textuales, en la cual se describen los fenómenos que pueden darse en el texto; y, finalmente, la documentación de los conjuntos de etiquetas empleados para describirlos.

En cuanto a la DTD principal, si bien es común a todos los textos codificados conforme a la TEI, permite ser usada de múltiples maneras según las necesidades impuestas por el texto y por el codificador. La variación radica en la posibilidad de combinar diversos conjuntos de etiquetas según lo que los editores de la TEI llaman jocosamente *"Chicago pizza model"*: "En Chicago, como en el resto de Estados Unidos (aunque no en Italia), todas las pizzas tienen algunos ingredientes en común (queso y salsa de tomate); no obstante, el consumidor puede especificar su opción por un tipo de masa (fina, gruesa, rellena) y una selección arbitraria de ingredientes adicionales [...]".[15] La analogía se basa en el hecho de que el codificador de documentos TEI puede adaptar

15. Sperberg-McQueen & Burnard (1995), p. 27.

la DTD principal combinando varios conjuntos de etiquetas disponibles en ella: un conjunto nuclear (que siempre está presente), un conjunto básico determinado y una selección cualquiera de conjuntos adicionales.

El usuario debe, pues, elegir un conjunto básico de etiquetas entre ocho disponibles, seis de ellos diseñados para documentos en los que predomina un tipo textual determinado (prosa, poesía, teatro, transcripción de material hablado, diccionarios impresos y datos terminológicos) y, los dos restantes, para documentos en los que se combinan distintos tipos textuales (una "base general" para antologías y una "base mixta" para combinaciones anárquicas de tipos textuales).

Además, el usuario puede seleccionar cualquier combinación de una serie de conjuntos adicionales de etiquetas, en los que se definen elementos relacionados con determinados tipos de procesamiento del texto, con determinados tipos de investigación, o bien elementos que pueden aparecer en diferentes tipos de texto pero cuyo uso menos generalizado no justifica su inclusión en el conjunto de etiquetas nuclear. Los conjuntos adicionales definen las etiquetas necesarias para representar enlaces hipertextuales, segmentaciones y alineaciones de textos, análisis e interpretaciones textuales, indicaciones sobre el grado de certeza en la codificación y sobre la responsabilidad de la misma, transcripciones de manuscritos, críticas textuales, análisis detallados de nombres o fechas, estructuras como grafos, redes, árboles, tablas, fórmulas o gráficos, información demográfica sobre los autores o hablantes e indicaciones de tipología textual.

La selección del conjunto de etiquetas básico y la de los adicionales, así como cualquier modificación de las definiciones propuestas por la TEI (como los cambios en los nombres de los elementos, o la adición de elementos nuevos), se realizan mediante marcas declarativas similares a las que hemos visto en nuestra introducción a SGML.

Concluiremos esta breve introducción al esquema de codificación TEI con una enumeración ilustrativa de algunos de los elementos definidos de uso más corriente en un documento TEI: por una parte, los que constituyen el conjunto de etiquetas nuclear y, por otra, los incluidos por defecto en todos los conjuntos de etiquetas básicos. En el conjunto de etiquetas nuclear, se definen elementos de tres tipos: i) los que pueden aparecer como componentes directos de una de las grandes divisiones de un texto; ii) los que pueden aparecer en los niveles de los caracteres individuales, de las palabras o de los fragmentos de texto inferiores a los componentes del primer tipo; iii) los que pueden aparecer tanto en el nivel de los componentes de divisiones textuales como en el de los fragmentos inferiores.

Los elementos definidos en el conjunto de etiquetas nuclear que pueden ser componentes directos de las divisiones textuales son los párrafos (etiqueta `<p>`), los versos (etiqueta `<l>`, de "verse line"), las estrofas u otros grupos de versos (`<lg>`, de "line group", y que contienen uno o más elementos `<l>`) y las intervenciones habladas de una obra teatral (`<sp>`, de "speech", y que pueden contener a su vez, además de una indicación opcional del hablante, `<speaker>`, una serie de párrafos, versos o grupos de versos).

Los elementos definidos en el conjunto nuclear que pueden aparecer en los niveles inferiores son mucho más variados por lo cual no los enumeraremos de forma exhaustiva. Incluyen, entre otros: fragmentos destacados tipográficamente (<hi>, de "highlighted"), con distinción opcional entre énfasis (<emph>), términos técnicos (<term>), palabras extranjeras (<foreign>), expresiones citadas (<mentioned>), etc.; elementos diferenciados semánticamente que puede resultar conveniente identificar, como nombres (<name>), números y medidas (<num> y <measure>), fechas y horas (<date>, <time>), etc.; modificaciones editoriales, como correcciones de errores (<corr>), reproducción consciente de errores (<sic>), añadidos (<add>, "addition"), texto transcrito a pesar de estar tachado en el original (<del>, "deleted"), etc.; y distintos elementos para la señalización de estructuras hipertextuales como referencias cruzadas y palabras indexadas.

Por último, los elementos que pueden aparecer tanto en el nivel de los componentes directos de las divisiones textuales como en niveles inferiores comprenden entre otros: las anotaciones (<note>), con la posibilidad de indicar, mediante el uso de atributos, su tipo y localización (notas al pie, finales, marginales); las listas y sus elementos (<list>, <item>); las citas textuales (<q>, de "quotation"); las citas bibliográficas (<bibl>); las acotaciones de escena de textos teatrales (<stage>); y los textos anidados (<text>).

En cuanto a los elementos incluidos por defecto en los ocho conjuntos de etiquetas básicos que hemos mencionado, se trata de aquellos que definen la estructura general de un documento. La estructura proporcionada como opción por defecto divide cualquier texto en tres unidades de máximo nivel: un cuerpo (<body>), que puede ir precedido opcionalmente por materia preliminar (<front>) y seguido, también opcionalmente, por materia posterior (<back>).

Estas grandes unidades suelen estar divididas en componentes que reciben nombres muy diversos según el tipo de texto (capítulos, secciones, subsecciones, actos, escenas, entradas, partes, libros, cantos, etc.). Todos ellos suelen adoptar estructuras arbóreas, en las que los componentes más pequeños se anidan dentro de otros de rango inmediatamente superior. Por ello, todos son tratados por el esquema de codificación de la TEI como un solo tipo de elemento llamado división textual ( ). Las etiquetas pueden incluir un atributo "type" para distinguir el nombre asociado con cada división concreta (por ejemplo,). La TEI ofrece también la posibilidad de optar por una serie de ocho elementos distintos para ocho niveles jerárquicos de división (<div0>, <div1>, ... <div7>). También se puede emplear simplemente el elemento para todos ellos, ya que la estructura jerárquica está implícita en el propio anidamiento de unos elementos dentro de otros superiores. Sin embargo, esto puede dificultar la detección de errores de codificación.

Cada división de un texto puede incluir un título al comienzo, el cual se codifica mediante una etiqueta específica (<head>). También hay otros elementos típicos de la estructura lógica de ciertos documentos, que suelen aparecer al principio o al final de una división determinada, como epígrafes, fechas, etc., para los cuales se definen también elementos específicos, que suelen incluirse dentro de otros dos elementos que los agrupan al principio (<opener>) o al final (<closer>) de la división en cuestión.

EAGLES: Hacia unas pautas comunes para la ingeniería lingüística

Hemos mencionado ya la creciente necesidad de contar con recursos lingüísticos en soporte informático tanto para el estudio de la lengua como para el desarrollo de productos de tecnología lingüística. La creación de recursos tales como corpus de textos escritos o de lenguaje hablado mediante la recopilación de materiales procedentes de fuentes diversas, su codificación en soporte informático, su anotación o enriquecimiento con información lingüística y el desarrollo de las herramientas necesarias para su gestión y explotación requieren un esfuerzo y una inversión de medios considerables. Lo mismo puede decirse de la compilación y mantenimiento de diccionarios en soporte informático, sea para uso humano o automático, o del desarrollo y perfeccionamiento de conjuntos de reglas gramaticales o de algoritmos estocásticos que proporcionen el conocimiento lingüístico necesario en cualquier sistema destinado a manipular el lenguaje natural. Resulta, por tanto, primordial que el diseño de estos recursos (corpus, diccionarios, gramáticas y herramientas especializadas en la manipulación de información lingüística) no responda a un tratamiento de los fenómenos lingüísticos pensado específicamente para aplicaciones determinadas. Se plantea, por el contrario, una necesidad de normalización, o al menos de armonización, de la información lingüística y de su representación formal.

Si bien este campo evoluciona a un ritmo que impide la fijación de estándares claramente establecidos, los avances realizados en él parecen indicar que sí es posible alcanzar un consenso para la elaboración de recomendaciones sobre determinados aspectos del desarrollo, la explotación y la evaluación de recursos lingüísticos. Para hacer frente a estas necesidades de normalización, y en el marco del programa comunitario de Investigación e Ingeniería Lingüística (LRE), la Dirección General XIII de la Comisión Europea puso en marcha en 1993 el proyecto *EAGLES (Expert Advisory Group on Language Engineering Standards)*. El objetivo principal de EAGLES es elaborar, mediante un amplio consenso, recomendaciones y especificaciones para áreas concretas de la tecnología lingüística a partir de los resultados de trabajos en curso en diversas organizaciones del ámbito comunitario y promover su adopción en futuros proyectos.

La puesta en marcha de EAGLES ha sido posible gracias al compromiso asumido por expertos de más de 30 centros de investigación, empresas, consorcios y asociaciones profesionales de la CE de aportar su tiempo y esfuerzo al trabajo del Grupo. El proyecto está coordinado por el Instituto de Lingüística Computacional de Pisa, y en su Consejo de Administración están representadas diversas empresas e instituciones académicas europeas, además de varias asociaciones y organismos de coordinación, también de ámbito europeo, como la Red Europea de Centros de Excelencia en Lenguaje y Habla (ELSNET), el capítulo europeo de la Asociación para la Lingüística Computacional (EACL), la Asociación Europea para la Comunicación del Habla (ESCA) y la Asociación Europea para la Lógica, el Lenguaje y la Información (FOLLI).

EAGLES pretende alcanzar un consenso de manera pragmática y realista para aquellas cuestiones que se presten a ello y señalar futuras direcciones de trabajo en los aspectos para los cuales tal consenso aún no es factible. Los resultados deben ser aplicables en la práctica al desarrollo de productos y

sistemas de procesamiento del lenguaje, y tan exhaustivos como lo permitan los avances técnicos en este campo. Además, deben ser compatibles con los estándares y recomendaciones existentes para campos tecnológicos afines, aplicables a distintas lenguas, y suficientemente abiertos y flexibles como para adaptarse a nuevos avances y a las distintas necesidades y puntos de vista de los usuarios.

La labor de definición de las especificaciones y recomendaciones está siendo llevada a cabo por cinco grupos de trabajo formados por expertos de empresas y universidades europeas, cada uno de los cuales se ocupa de una de las siguientes áreas: corpus textuales, léxicos computacionales, formalismos gramaticales, evaluación de productos de tecnología lingüística y lengua hablada. Dado que el trabajo de elaboración de las primeras recomendaciones está actualmente en marcha,[16] nos limitaremos a mencionar las principales tareas abordadas por cada uno de estos grupos. Comenzaremos por el Grupo de *Corpus Textuales*, con el cual el Instituto Cervantes ha colaborado en calidad de sede técnica y administrativa.

La normalización en la creación y explotación de corpus lingüísticos requiere, en primer lugar, la definición de un conjunto de parámetros para la clasificación y tipificación de corpus y de textos, ya que, para que un corpus sea realmente útil, es imprescindible que tanto los textos que contiene como el propio corpus puedan ser clasificados dentro de una tipología clara. Por otra parte, se ha tratado la elaboración de una solución especializada para hacer frente a las necesidades específicas de la codificación de corpus lingüísticos a partir de las recomendaciones de la TEI.

En lo referente a las normas de anotación lingüística de corpus, puesto que una normalización demasiado rígida de los sistemas de anotación morfosintáctica o etiquetado no resulta recomendable debido a las diferentes necesidades de cada proyecto, el trabajo de EAGLES se orienta hacia un marco general que permita diseñar esquemas concretos de anotación que sean compatibles. No obstante, además de este marco general, se proponen especificaciones por defecto que pueden ser adoptadas cuando no haya motivos especiales que requieran el desarrollo de esquemas específicos. Este marco general es susceptible de ser extendido para lograr la cobertura de fenómenos específicos de determinada lengua o lenguas. Además, permite la adopción de diversos grados de granularidad en la anotación. Por otra parte, se están elaborando recomendaciones preliminares para la anotación sintáctica, en especial para la descripción de información sintáctica superficial, como es el caso de la delimitación de sintagmas.

Otros asuntos que han sido objeto de estudio por parte del Grupo de Corpus Textuales han sido la documentación que debe acompañar a un corpus, la definición de un entorno de usuario y de las herramientas de que debe constar, el tratamiento de los corpus de textos paralelos (es decir, textos producidos en

16. No obstante, algunos de los documentos preparados por los grupos de trabajo están ya disponibles en Internet (*vid.* Bibliografía).

varios idiomas o traducidos a varios idiomas), y los métodos de transcripción y representación del lenguaje hablado.

A continuación, describiremos de forma somera algunos de los cometidos de los restantes Grupos de Trabajo.

El Grupo de Trabajo sobre *Léxicos Computacionales* tiene como principal objetivo la elaboración de estándares para recursos léxicos reutilizables para procesamiento del lenguaje natural. Dichos estándares se están elaborando principalmente a partir de la evaluación e integración de resultados ya alcanzados en diversos proyectos. Los trabajos del Grupo se ocupan de los distintos niveles de información lingüística presentes en los mismos, especialmente en el nivel morfosintáctico, para el cual se ha trabajado en estrecha colaboración con el subgrupo encargado de anotación morfosintáctica de corpus, adoptando el mismo enfoque en lo que atañe a la flexibilidad en el grado de granularidad elegido y en la posibilidad de extender el esquema para cubrir fenómenos específicos. Para el nivel sintáctico, se ha realizado un estudio comparativo de cómo se abordan los fenómenos de subcategorización en diferentes sistemas y teorías para diferentes lenguas europeas, y se ha elaborado un esquema preliminar de clasificación, también en colaboración con el correspondiente subgrupo de anotación de corpus. El Grupo de Trabajo sobre *Formalismos Gramaticales* se ha propuesto alcanzar un consenso sobre las características básicas de los formalismos aplicados al procesamiento del lenguaje natural e indicar posibles tendencias y necesidades futuras a partir de un estudio exhaustivo de los formalismos existentes y de las necesidades planteadas por la industria, con especial atención a las novedades en materia de descripción gramatical y desarrollo de gramáticas. El Grupo sobre *Evaluación* ha desarrollado un marco general para el diseño de evaluaciones. El objetivo es mejorar los métodos de evaluación, como primer paso hacia el establecimiento de estándares para productos de tecnología lingüística, así como identificar y especificar los componentes de un compendio de criterios de evaluación y técnicas asociadas, junto con recomendaciones sobre su uso, del cual los profesionales encargados de la evaluación puedan seleccionar las técnicas adecuadas para sus propósitos. El Grupo de Trabajo sobre *Lengua Hablada* se ha ocupado de las especificaciones aplicables a la producción de corpus de lengua hablada, así como a la evaluación de los sistemas de conversión de texto escrito a lengua hablada y viceversa, y otras tecnologías de esta área, como identificación del hablante e identificación de la lengua.

En el momento de preparar este artículo, se prevé la publicación inminente de un conjunto de especificaciones y recomendaciones que incluirán, para cada una de las áreas tratadas, un estudio sobre el estado actual de la cuestión, y una evaluación del grado de adecuación de las soluciones existentes para ser tenidas en cuenta como modelos genéricos de sus dominios respectivos. Para aquellas áreas cuyo grado de desarrollo ha permitido alcanzar un consenso, se incluirán recomendaciones basadas en dicha evaluación, o propuestas que complementen o modifiquen las soluciones ya existentes. Se identificarán asimismo los puntos que no hayan podido ser tratados debidamente en el curso del proyecto y se propondrán acciones futuras al respecto.

Referencias bibliográficas

BURNARD, L. (1995) "What Is SGML and How Does It Help?", *Computers and the Humanities*, 29, 41-50.(*)

CALZOLARI, N., BAKER, M., KRUYT, T. (eds.) (1994) «Towards a Network of European Reference Corpora», report of the *NERC Consortium Feasibility Study*. Pisa: Giardini Editori e Stampatori.

EAGLES, documentos de trabajo: En el momento de preparar este artículo, estaba prevista la publicación inminente de una primera serie de documentos de trabajo de EAGLES. Algunos de ellos estaban ya disponibles para su consulta a través de Internet en la siguiente dirección: "http://www.ilc.pi.cnr.it/EAGLES/home.html".

GOLDFARB, C.F. (1990) *The SGML Handbook*. Oxford: Oxford University Press.

HEID, U., MCNAUGHT, J. (eds.) (1991) *EUROTRA-7 Study: Feasibility and project definition study of the reusability of lexical and terminological resources in computerised applications*. Final report submitted to the CEC. Stuttgart: IMS, University of Stuttgart.

IDE, N.M., SPERBERG-MCQUEEN, C. M. (1995) "The TEI: History, Goals and Future", *Computers and the Humanities*, 29, 5-15. (*)

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO) (1986) ISO 8879: *Information Processing - Text and Office Systems - Standard Generalized Markup Language (SGML)*. Geneva: ISO.

ROTHENBERG, J. (1995) "Ensuring the Longevity of Digital Documents", *Scientific American*, January, 24-29.

SPERBERG-MCQUEEN, C. M., BURNARD, L. (1994) *Guidelines for Electronic Text Encoding and Interchange*. Chicago and Oxford: ACH-ACL-ALLC Text Coding Initiative.

— (1995) "The Design of the TEI Encoding Scheme", *Computers and the Humanities*, 29, 17-39. (*)

(*). Estos tres artículos, así como otros sobre aspectos más concretos de la TEI, constituyen el contenido de un número monográfico de *Computers and the Humanities* y se reproducen también en IDE, N. y J. VÉRONIS (eds.), 1995. *Text Encoding Initiative: Background and Context*. Dordrecht: Kluwer Academic Publishers.