VOLUMEN MONOGRÁFICO (1999), 25-52

# ESTÁNDARES DE ANOTACIÓN EN LINGÜÍSTICA DE CORPUS[1]

JAVIER PÉREZ GUERRA
*Universidad de Vigo*

RESUMEN. *La universalización de los ordenadores en la comunidad investigadora ha conducido al desarrollo de nuevas metodologías basadas en el acceso inmediato a datos lingüísticos reales pertenecientes a colecciones electrónicas de textos de grandes dimensiones. El auge de estos estudios empíricos ha dado lugar al nacimiento de una nueva disciplina, la llamada Lingüística de Corpus. En este trabajo se introduce el concepto de 'corpus de textos' y las consecuencias teóricas de su utilización dentro de los paradigmas lingüísticos habituales. La parte central de este estudio presta atención especial a la inclusión de información no textual en un corpus de textos, esto es, a su 'anotación'. Después de recorrer el espectro de diferentes esquemas de anotación lingüística, nos detendremos en el estándar SGML-TEI.*

PALABRAS CLAVE. *Lingüística de Corpus, textos electrónicos, anotación textual, SGML, TEI.*

ABSTRACT. *The generalisation of computers in the research community has led to the development of new methological approaches based on the immediate access to actual linguistic data which are included in huge eletronic collections of texts. The success of such empirical studies has paved the way for the birth of a new science, namely, so-called Corpus Linguistics. In this paper 1 introduce the concept of 'corpus of texts' as well as the theoretical consequences of its use as far as the most popular linguistic models are concerned. In the main sections of this paper special attention is paid to the inclusion of non-textual information in textual corpora, that is, to the 'annotation' of corpora. Once the spectrum of the various schemes of linguistic annotation has been explored, 1 shall focus on the standard mark-up approach SGML-TEI.*

KEYWORDS. *Corpus Linguistics, electronic texts, textual annotation, SGML, TEI.*

## 1. INTRODUCCIÓN[2]

Las alternativas de modelización teórica del estudio de las lenguas pueden reducirse a dos opciones claramente diferenciadas. Por un lado, el foco de la explicación lingüística puede ser la lengua real tal y como la percibimos bien en el

medio escrito bien en el medio oral. Por otro, los lingüistas pueden estar más interesados en describir y justificar la capacidad interna de los seres humanos a la hora de "producir" actos lingüísticos mediante fórmulas abstractas complejas que sirven como puente entre la producción real (una cuestión de *performance*) y el diseño mental filosófico que se esconde detrás de las producciones lingüísticas (la llamada *competence*) — sobre los conceptos de *competence* y *performance*, véase, por ejemplo, Chomsky (1965: 44ss). La primera alternativa, generalmente conocida como "funcional", selecciona la lengua real como punto de partida con el fin de explicar la organización, variación y producción (de nuevo) "funcionales" de cada uno de los elementos que la componen.

Aunque la finalidad de los dos modelos teóricos introducidos en el párrafo anterior coincide, ya que ambos pretenden la explicación de actos lingüísticos aceptables, existen diferencias sustanciales tanto en la determinación de lo que es "aceptable" en una lengua como en la metodología que los lingüistas siguen a la hora de justificar dichos actos (véase Stubbs 1996: capítulo 2). Por lo que respecta al concepto de aceptabilidad, en opinión de los funcionalistas una expresión lingüística es aceptable en un código lingüístico (y, en consecuencia, idónea para ser estudiada en términos lingüísticos) si ha sido (o puede ser) producida por los hablantes de dicha lengua. Por el contrario, la noción de aceptabilidad en escuelas formales reside en los juicios que los hablantes individuales de una lengua poseen sobre una determinada producción lingüística. Desde el punto de vista de la metodología, mientras que las escuelas formales utilizan el que bien pudiéramos llamar "método de despacho" (véase Fillmore 1992: 35), esto es, la introspección reflexiva, una buena parte de los funcionalistas (posiblemente, menos de lo deseable) basan sus estudios en trabajos de campo, es decir, en la prospección externa.

En este trabajo nos centraremos en abordar cómo la informática puede colaborar a la hora de llevar a buen puerto el primero de los niveles de la investigación que tenemos entre manos, esto es, la obtención de datos reales. Surge así el concepto de "corpus electrónico de textos" como una realidad manipulable mecánicamente, de la cual el lingüista puede extraer rápida y fiablemente información real sobre la lengua.

## 2. UNAS PINCELADAS HISTÓRICAS SOBRE LA LINGÜÍSTICA DE CORPUS

El uso de colecciones de textos en la investigación lingüística no debe ser considerado algo novedoso. Si bien el análisis lingüístico computarizado ya había sido iniciado a finales de los años 40 por el Padre Busa con su concordancia de la obra de Santo Tomás de Aquino (Busa 1974-80), podríamos tomar la década de los 60 (o incluso antes) como fecha inicial de una primitiva Lingüística de Corpus. Los primeros corpus de textos, consistentes en colecciones de material textual tomado de libros, periódicos, televisión, conversaciones y similares, se utilizaban principalmente

como base empírica que se añadía a tests realizados ante hablantes nativos de una lengua. Es a finales de los años 60 cuando aparece un corpus de textos tal y como lo entendemos hoy: el popular *Survey of English Usage* (*SEU*), que consistía en un gran archivo de fichas de cartón con material oral transcrito por un grupo de investigadores dirigidos por Randolph Quirk. A partir de este momento surgen tres corpus de textos: el *Brown University Corpus* en 1964 (un millón de palabras de textos publicados en 1961 en los Estados Unidos de Norteamérica) bajo el liderazgo de N. Francis y H. Kucera [la versión anotada fue publicada en 1979], el *Lancaster-Oslo/Bergen Corpus* o *LOB* en 1978 (versión británica del *Brown Corpus*) [versión etiquetada en 1986] y el *London-Lund Corpus of Spoken English* de 1980 (versión electrónica del *Survey* llevada a cabo por el equipo de Jan Svartvik). Estos tres nuevos corpus contenían textos en formato electrónico, gracias a las ventajas que ofrecía una ya asequible informatización de textos en ordenadores. Sin duda, la aparición en el mercado académico de estas tres colecciones iniciaría lo que se conoce como la Lingüística de Corpus "D.C", a saber, 'Después de los Computadores'.

Aquellas primeras técnicas de selección, clasificación, informatización y anotación aplicadas a estos corpus estaban todavía lejos de proyectos ya disponibles del alcance del *Collins Birmingham University International Language Database* (*COBUILD*) (cerca de 20 millones de palabras de inglés escrito y oral) [más información en *http://titania.cobuild. collins.co.uk/boe_info.html*, desde donde podemos realizar una búsqueda de demostración] o el *British National Corpus* (*BNC*) (cerca de 100 millones de palabras de inglés británico hablado y escrito) [más información en *http://info.ox.ac.uk/bnc*; acceso en línea desde *http://thetis.bl.uk/lookup.html*], este último iniciado en 1991 bajo la protección de Oxford University Press y aún no ultimado de modo definitivo.

Coincidiendo con el auge del más puro empiricismo académico fue cuando Noam Chomsky, principal representante de la tradición formalista, cambió el rumbo de la Lingüística hacia un mayor racionalismo. En otras palabras, se pasó de la observación de datos naturales a la introspección a la que hacíamos alusión más arriba. Como consecuencia de su posicionamiento teórico, tanto Chomsky como sus seguidores rechazaron el uso del corpus de textos como única fuente proporcionadora de datos. Para ellos, todo corpus de textos es parcial, en tanto en cuanto retrata exclusivamente actos de *performance* y no de *competence*. Para esta escuela de la Lingüística, es el hablante (o el lingüista) quien mejor puede determinar su *competence*. En palabras de McEnery y Wilson (1996: 9), según Chomsky, un hablante nativo de una lengua es "the sole explicandum of linguistics".

La enorme expansión de las ideas de Chomsky no provocó, sin embargo, el abandono completo de la investigación mediante corpus de textos. Incluso hoy en día no resulta extraño del todo encontrar estudios situados teóricamente en modelos formales que hacen uso de corpus estándar de textos a la hora de probar o rechazar una determinada hipótesis.

## 3. CONCEPTO DE "CORPUS DE TEXTOS". HACIA EL CORPUS ELECTRÓNICO

Un "corpus de textos" es simplemente, en palabras de Johansson (1991: 3), "a body of texts put together in a principled way, often for the purposes of linguistic research", esto es, un conjunto de textos reales y, de ahí, aceptables pertenecientes a un código lingüístico determinado.

Además del de ser un conjunto de textos, la acepción de "corpus" que aquí vamos a utilizar ha de cumplir un segundo requisito, a saber, su manejabilidad mediante procedimientos electrónicos. Sirva la definición de Sánchez *et al.* (1995: 8-9) como apoyo teórico para este uso restringido de "corpus":

> Un corpus lingüístico es un conjunto de datos lingüísticos (pertenecientes al uso oral o escrito de la lengua, o ambos) sistematizados según determinados criterios, suficientemente extensos en amplitud y profundidad de manera que sean representativos del total del uso lingüístico o de algunos de sus ámbitos, y dispuestos de tal modo que puedan ser procesados mediante ordenador con el fin de obtener resultados varios y útiles para la descripción y análisis.

Aquellos lingüistas que requieran evidencias reales de un acto lingüístico concreto encontrarán que estas colecciones de textos constituyen un medio extremadamente útil a la hora de iniciar sus investigaciones, pues la mayoría de las posibilidades que ofrecen estos corpus textuales no pueden conseguirse mediante la introspección individual. De entre sus muchos méritos, citaremos la determinación de la frecuencia de una palabra en un texto o grupo de textos, de todas las apariciones de una palabra en un contexto específico o con una función o categoría gramatical concreta, de las relaciones léxicas entre grupos de palabras, esto es, las llamadas "colocaciones", de la longitud (número de palabras) y complejidad de las oraciones de los diferentes tipos de textos, de cuestiones de variación sintáctica (concordancia, orden de palabras), etc.

No debemos, sin embargo, pensar que todos los corpus son infalibles. Un buen número de factores están directamente relacionados con el grado de éxito de un determinado corpus. Nos referiremos en lo que sigue a tres de ellos: el tamaño del corpus, el rigor en la selección del material y la codificación de los textos.

Cuanto mayor sea el volumen de los textos recogidos, más oportunidades tendremos bien de encontrar la secuencia buscada o bien de desterrarla con cierta seguridad de que no es aceptable. Ciertamente, el corpus perfecto sería aquel que incluyese la totalidad de las secuencias lingüísticas aceptables y realizables de una lengua. Obviamente, al tratarse de corpus de lenguajes naturales, la recopilación de todas las producciones lingüísticas no es posible, por lo que nunca podremos alcanzar una fiabilidad del cien por cien. Es en este estado de cosas en el que, en nuestra opinión, debe entrar en juego precisamente tanto el elemento introspectivo como conceptos más estadísticos del tipo de la "probabilidad" de aparición de una secuencia, etc.

La inclusión de un número enorme de textos dificulta tremendamente el manejo del corpus, de ahí que en muchos casos tengamos que proceder a la formalización de estos en un formato que pueda ser entendido por una máquina. Desde esta perspectiva, lo que hace no demasiados años se consideraba un "gran" corpus textual (un millón de palabras) resulta prácticamente ridículo si lo comparamos con los estándares que se están manejando en la actualidad (proyectos como el *BNC* o el *Bank of English* de *COBUILD* ya reseñados). Numerosos factores contribuyen al aumento medio del tamaño de los corpus. Por un lado, los avances técnicos en los equipos y software de ROC (Reconocimiento óptico de Caracteres) facilitan el escaneado de textos escritos. Por otro lado, hoy resulta relativamente sencillo encontrar textos ya disponibles en formato electrónico, es decir, preparados para ser introducidos directamente en un ordenador. Prácticamente todos los trabajos de autoedición profesional (universidades, periódicos, editoriales, etc.) parten ya de textos electrónicos y no de versiones "en papel". Internet ha contribuido enormemente a la proliferación de corpus informatizados, dada la ingente cantidad de texto distribuida a través de la "telaraña mundial". Finalmente, la calidad de los equipos informáticos ha aumentado de manera inversamente proporcional a su precio, lo cual permite que cualquier usuario, centro de investigación, etc. pueda disponer de equipos para almacenar corpus textuales que hace unos años hubieran necesitado plataformas específicas y prohibitivas.

En cuanto al rigor de selección de los materiales, los textos que conforman un corpus deben reunir aquellas condiciones que aseguren su representatividad[3]. Surge de este precepto la necesidad de distinguir entre corpus de alta versatilidad, esto es, colecciones electrónicas de gran magnitud que albergan una alta variedad de registros, temas, voces, estilos, géneros, etc., y corpus específicos diseñados para la investigación lingüística de un tema concreto en unas circunstancias muy determinadas. Entre los corpus versátiles, podemos citar el *BNC* o el *Bank of English* en el caso de la lengua inglesa, o, para el español, el *Archivo de Textos Hispánicos de la Universidade de Santiago de Compostela (ARTHUS)* [más información en *http://www.usc.es/~sintx/Arthus.html],* el *Corpus oral de referencia del español contemporáneo* (Universidad Autónoma de Madrid) [más información en *http://lola. lllf. uam. es/docs_es/corpus/corpus.html] y CUMBRE,* desarrollado por SGEL y coordinado desde la Universidad de Murcia (véase Sánchez *et al.* 1995). Por lo que respecta al *ARTHUS,* la parte contemporánea comprende en el momento de escribir estas líneas 34 textos narrativos, teatrales, periodísticos y orales procedentes de España e Hispanoamérica, con un total de aproximadamente 1.500.000 palabras (todo el corpus recoge cerca de 3 millones de palabras). Por otro lado, la parte medieval contiene 13 textos con más de un millón de formas. En segundo lugar, la base de datos textual del *Corpus oral* incluye la transcripción de más de un millón de palabras de lengua española hablada tomada de situaciones diversas (intervenciones de los campos administrativo, científico, conversacional o familiar, educativo, humanístico, jurídico, lúdico, político, periodístico). Finalmente, *CUMBRE,* que

aspira a tener unos 30 millones de palabras, ya posee al menos 8 millones en formato electrónico. La Real Academia Española, a través de su Instituto de Lexicografía, trabaja en dos proyectos enmarcados en la confección de corpus lingüísticos del español: el *Corpus de Referencia del Español Actual* (*CREA*) y el *Corpus Diacrónico del Español* (*CORDE*). El primero, que abarca textos y transcripciones de español oral desde 1975 al año 2000, comprende ya más de 100 millones de palabras, lo que equivale a aproximadamente la mitad de la extensión final proyectada. *CORDE*, por otro lado, se centra en ejemplos de la lengua española desde sus orígenes hasta 1975. Unos 70 millones de palabras han sido adaptadas al medio electrónico [más información tanto de estos proyectos como de otros más específicos puede obtenerse del documento "Informe sobre recursos lingüísticos para el español (II): Corpus orales y escritos disponibles y en desarrollo en España", *http://www.cervantes.es/oeil/Oeiltipo.htm*].

Con la premisa de que no todos los corpus son válidos para todos los fines, el investigador puede estar interesado en estudios sintácticos, fonológicos, pragmáticos o semánticos, además de los puramente léxicos. Un corpus que contenga únicamente información léxica, esto es, palabras, poco puede ayudar en el terreno de la investigación, por ejemplo, sintáctica o fonológica. Es por ello por lo que dentro de la llamada Lingüística de Corpus hay fuertes líneas de investigación dirigidas hacia la anotación sintáctica, fonológica, fonética, semántica o discursivo-pragmática, con especial énfasis en la automatización de esos procesos. Una vez introducidos los conceptos básicos (véase igualmente Pérez *et al.* en este volumen) en las restantes páginas nos centraremos en el examen de distintos tipos de anotación o etiquetación desarrollados mayoritariamente en corpus populares de textos ingleses con el fin de efectuar una revisión de los estándares de incorporación de información lingüística a las colecciones electrónicas de textos.

## 4. ANOTACIONES EN LOS CORPUS

Habíamos ya indicado en la sección anterior que las posibilidades de análisis textual de corpus formados por exclusivamente texto se reducen a la exploración de fenómenos léxicos y/o tipográficos. Es más, esos resultados no pueden ser filtrados por otras variables, pues las herramientas de análisis textual aplicadas a textos puros tienen que limitarse a búsquedas de secuencias alfanuméricas y no pueden encadenar los resultados de dichas operaciones con información adicional de la que carecen estos corpus. Con el fin de ampliar el espectro de las posibilidades de análisis textual, surge el texto anotado o etiquetado, que incluye, además de texto puro, información adicional relativa a varios aspectos relacionados bien específicamente con las pretensiones investigadoras del anotador o bien con las necesidades generales de un sector de la comunidad investigadora.

Realizar una tipología de posibles sistemas de anotación no es un cometido sencillo, pues no existe un estándar ni de lo que se quiere anotar ni de cómo esto se ha de llevar a cabo. Por ello, en esta sección nos limitaremos a ofrecer un panorama de las categorías tradicionales de anotación, así como de los distintos sistemas que se han empleado (§4.1). En relación con este último aspecto, en §4.2 nos centraremos en el estándar de etiquetado SGML-TEI o *Standard[ized] General Markup Language — Text Encoding Initiative.*

## 4.1. *Categorías de anotación*

En esta sección ejemplificaremos un buen número de elementos de anotación que comúnmente se añaden a los textos crudos.

### 4.1.1. *Anotación textual*

Cuando adaptamos al formato electrónico material textual, podemos optar por transferir exclusivamente los textos sin incorporar más información que la meramente necesaria para identificar el pasaje. En el siguiente ejemplo, tomado del *Melbourne-Surrey Corpus*, un corpus de aproximadamente 100.000 palabras procedentes de textos de periódicos australianos de los años 1980 y 1981, comprobamos que las anotaciones se limitan a la codificación de cada texto — <F1> posiblemente indique que este es el texto o archivo número uno:

<F1>
THE AGE MONDAY 1 SEPTEMBER 1980
EDITORIAL OPINION
"A BLAND BUT USEFUL START"
THE COMMITTEE OF INQUIRY INTO THE AUSTRALIAN
FINANCIAL SYSTEM HAS RELEASED ITS INTERIM
REPORT AND ON FIRST READING IT IS A DISAPPOINTING
DOCUMENT. THE 572-PAGE REPORT CONTAINS NO
RECOMMENDATIONS, NO COMMENT ON THE MAJOR
ISSUES RAISED AND LITTLE ANALYSIS OF THOSE ISSUES

Cuadro 1: *Melbourne-Surrey Corpus*

La versión anotada del *LOB* del Cuadro 2 incorpora además información sobre la tipografía y la distribución gráfica del texto:

```
A01   1 **[001 TEXT A01**]
A01   2 *<*'*7STOP ELECTING LIFE PEERS**'*>
A01   3 *<*4By TREVOR WILLIAMS*>
A01   4  |^A *0MOVE to stop \0Mr. Gaitskell from nominating any more Labour
A01   5 life Peers is to be made at a meeting of Labour {0M P}s tomorrow.
A01   6  |^\0Mr. Michael Foot has put down a resolution on the subject and
A01   7 he is to be backed by \0Mr. Will Griffiths, {0M P} for Manchester
A01   8 Exchange.
A01   9  |^Though they may gather some Left-wing support, a large majority
A01   10 of Labour {0M P}s are likely to turn down the Foot-Griffiths
A01   11 resolution.
```

Cuadro 2: *Versión anotada del LOB*

Mientras que 'A01' es un simple indicador del texto, y los números de la segunda columna cuantifican el número de la línea, otros elementos precisan el tipo de letra, la naturaleza léxica de las cadenas, etc. Así, aquel segmento delimitado por las marcas '*<' y '>*' es un titular tipográfico; '|' y '^' indican, respectivamente, el inicio de un párrafo y una oración; '*0' implica letra normal (tamaño de cuerpo de texto); las cadenas entre las llaves '{' y '}' no son propias de la lengua inglesa, etc.

### 4.1.2. *Anotación gramatical*

Además de aquellos etiquetados tipográficamente, existen corpus con anotación gramatical. Distinguiremos aquí aquellos con anotaciones categoriales (*tagged*) y aquellos con información sintáctica (*parsed*).

El mismo texto del *LOB* que empleábamos más arriba para ilustrar la anotación tipográfica nos servirá ahora para ejemplificar la anotación categorial:

```
A01   2   ^ *'_*' stop_VB electing_VBG life_NN peers_NNS **'_**' ._.
A01   3   ^ by_IN Trevor_NP Williams_NP ._.
A01   4   ^ a_AT move_NN to_TO stop_VB \0Mr_NPT Gaitskell_NP from_IN
A01   4   nominating_VBG any_DTI more_AP labour_NN
A01   5   life_NN peers_NNS is_BEZ to_TO be_BE made_VBN at_IN a_AT meeting_NN
A01   5   of_IN labour_NN \0MPs_NPTS tomorrow_NR ._.
A01   6   ^ \0Mr_NPT Michael_NP Foot_NP has_HVZ put_VBN down_RP a_AT
A01   6   resolution_NN on_IN the_ATI subject_NN and_CC
A01   7   he_PP3A is_BEZ to_TO be_BE backed_VBN by_IN \0Mr_NPT Will_NP
A01   7   Griffiths_NP ,_, \0MP_NPT for_IN Manchester_NP  A01   8 Exchange_NP ._.
A01   9   ^ though_CS they_PP3AS may_MD gather_VB some_DTI left-wing_JJB
A01   9   support_NN ,_, a_AT large_JJ majority_NN
```

| | |
|---|---|
| A01 10 | of_IN labour_NN \0MPs_NPTS are_BER likely_JJ to_TO iurn_VB down_RP |
| A01 10 | the_ATI Foot-Griffiths_NP |
| A01 11 | resolution_NN ._. |

Cuadro 3: *LOB con etiquetado morfosintáctico*

En esta reproducción, además de toda la información ya expresada en la versión no anotada categorialmente, comprobamos que cada palabra está asociada con algún tipo de etiqueta que permite adivinar su naturaleza gramatical. Así, por ejemplo, en la línea 2, *stop* y *electing* están caracterizadas como formas verbales ('VB' indica la forma base de un verbo y 'VBG' el participio de presente); *life* y *peers* como nombres comunes en singular y plural, respectivamente ('NN' y 'NNS'); *Trevor* y *Williams* en la línea 3 como nombres propios en singular ('NP'); *by* como preposición ('IN'), etc.

Los sistemas de anotación categorial son muchos y muy variados. Frente al conjunto de 152 etiquetas empleadas en la anotación del *London-Lund* o las 132 del *LOB*, el *Brown* emplea 87, 57 el *BNC* y sólo 25 el *Penn TreeBank*. Igualmente, el *modus operandi* del proceso de anotación puede ser totalmente manual, totalmente automático o mixto. De entre los sistemas automáticos de anotación categorial, el CLAWS ocupa un lugar relevante (véase Garside y Smith 1997). De hecho, el porcentaje de éxito de este sistema en la etiquetación del *BNC* ha sido del 97 por ciento, lo cual, en un corpus que contiene tanto material textual como el *BNC*, equivale a la fiabilidad absoluta. En la actualidad existen etiquetadores automáticos gratuitos, disponibles en la red Internet. Indicaremos a continuación las direcciones en las que podemos obtener más información: *Bill-Tagger* para Unix, DOS y Mac (*ftp://ftp.cs.jhu.edu/pub/brill/Programs/*), *Xerox-Tagger* (*http://www.xerox.fr/grenoble/mltt/home.html*), o *Birminham e-mail tagger* (*http://clg1.bham.ac.uk/tagger.html*).

Por lo que respecta a la anotación sintáctica o *parsing* –'sintáctica' aquí entendida como aquella información que supera el nivel de la palabra–, el número de corpus etiquetados según este criterio es inferior, pues la dificultad que conlleva la automatización del proceso es muy alta (sobre *parsing* –automático–, véase Leech y Garside 1991, Souter y O'Donoghue 1991, Qiao 1995 y Bunt y Tomita eds. 1996 entre otros). Manejaremos a continuación ejemplos del *Lancaster Parsed Corpus* (sección de aproximadamente 140.000 palabras del *LOB*), *SUSANNE Corpus* (130.000 palabras del *Brown*) y el *Penn TreeBank* (conjunto de textos electrónicos de diferentes extensiones, etiquetados categorial y sintácticamente por un equipo de investigadores de la Universidad de Pennsylvania).

```
A07 418
[S[Na I_PP1A Na] [V can_MD n't_XNOT make_VB V][N a_AT club_MM N][Tb[V
pay_VB V] [N a_AT player_NN N][N[D so_QL much_AP D][N a_AT week_NN N]N]
Tb]._. S]
B04 248
[S[N \OMr_NPT Henry_NP Newton_NP [Po of_INO [N Acton_NP N] Po]N][V does_DOZ
not_XNOT want_VB V][N his_PP$ daughter_NN N] [Ti [Vi to_TO marry_VB Vi][N
a_AT Scotsman_NNP N)Ti] ._. S]
```

Cuadro 4: *Lancaster Parsed Corpus*

En este ejemplo del *Lancaster* observamos el tipo de anotación que aquí tildamos de "sintáctica". Así, además de la información categorial correspondiente a cada una de las unidades léxicas del pasaje ('PP1A' etiqueta el pronombre personal de primera persona *I*, 'MD' se emplea con modales como *can*, 'XNOT' con partículas de negación —*not*—, 'VB' como verbos léxicos —*make*—, 'AT' con artículos como *a*, etc.), el anotador utiliza elementos que van más allá del nivel de la palabra. A modo de ejemplo, 'N' y 'Na' definen las frases nominales *I*, *a club*, *a player*, *Mr Henry Newton*; 'V' señala qué elementos forman parte del grupo verbal (*can't make*, *pay*, *does not want*, etc.); 'Po' delimita frases preposicionales como *of Acton*, etc. Y lo que es más, 'S' se emplea para marcar oraciones (*I can't make a club pay a player so much a week*, *Mr Newton of Acton does not want his daughter to marry a Scotsman*).

Un proyecto que hará las veces de prototipo en este recorrido por la tipología de corpus *parsed* o analizados sintácticamente es el *SUSANNE*. Se trata este de un corpus pequeño, cuya utilidad no es obviamente la de servir como elemento textual para prospecciones léxicas o gramaticales. Al contrario, *SUSANNE* se ha caracterizado por ser un mero proyecto de un complejo y, a la vez, completo sistema (o, como prefiere Geoffrey Sampson, "esquema") de análisis integral de, en este caso, la lengua inglesa. Nos limitaremos a ejemplificar en Cuadro 5 alguno de los potenciales del modo de anotación empleado en *SUSANNE* [adaptación gráfica del original]. Para más información, el lector deberá consultar directamente el manual "oficial" de este corpus (Sampson 1995):

```
N05:0010a -  PPHS1f   She        she        [O[çs[Nas:s.Nas:s]
N05:0010b -  VBDZ     was        be         [Vsu.
N05:0010c -  VVGv     carrying   carry      .Vsu]
N05:0010d -  AT1      a          a          [Ns:o.
N05:0010e -  NN1c     quirt      quirt      .Ns:0]
N05:0010f -  YC       +,         –          .
N05:0010g -  CC       and        and        [S+.
N05:0010h -  PPHS1f   she        she        [Nas:s.Nas:s]
```

| | | | | | |
|---|---|---|---|---|---|
| N05:0010i | - | VVDv | started | start | [Vd.Vd] |
| N05:0010j | - | TO | to | to | [Ti:z[Vi. |
| N05:0010k | - | VV0t | raise | raise | .Vi] |
| N05:0010m | - | PPH1 | it | it | [Ni:o.Ni:o]Ti:z] |
| N05:0010n | - | YC | +, | - | . |
| N05:0010p | - | RTn | then | then | [S-[Rsw:t.Rsw:t] |
| N05:0010q | - | VVDv | let | let | [Vd.Vd] |
| N05:0010r | - | PPH1 | it | it | [Ni:O102.Ni:O102] |
| N05:0010s | - | YG | - | - | [Tb:o[s102.s102] |
| N05:0010t | - | VV0i | fall | fall | [V.V] |
| N05:0020a | - | RT | again | again | [R:t.R:t] |
| N05:0020b | - | CC | and | and | [Tb+. |
| N05:0020c | - | VV0v | dangle | dangle | [V.V] |
| N05:0020d | - | II | from | from | [P:q. |
| N05:0020e | - | APPGf | her | her | [Ns. |
| N05:0020f | - | NN1c | wrist | wrist | .Ns]P:q]Tb+]Tb:o]S-]S+]S] |
| N05:0020g | - | YF | +. | - | .O] |
| N05:0020h | - | YB | <minbrk> | - | [Oh.Oh] |

Cuadro 5: *SUSANNE Corpus*

En el sistema de codificación de *SUSANNE*, cada palabra ocupa una línea, la cual está dividida en seis campos separados por espacios en blanco o tabulaciones. Así, la palabra real del texto está incluida en el cuarto campo. De hecho, si leemos la cuarta columna, podremos reconstruir el texto original: *She was carrying a quirt, and she started to raise it, then let it fall again and dangle from her wrist.* El primer campo se emplea como identificador de la posición de la palabra: los tres primeros caracteres constituyen el código correspondiente al texto en el *Brown* (N05 en este ejemplo); los cuatro caracteres que siguen al separador ':' se emplean como numeradores de las líneas en las que las palabras en cuestión aparecen en el texto original; finalmente, cada palabra en cada línea está identificada mediante una letra minúscula. El segundo campo está reservado para información léxica general sobre la palabra analizada. En nuestro ejemplo, este está cubierto con la opción por defecto, esto es, el guión '-'. El campo siguiente está ocupado por la etiqueta correspondiente a la categoría morfosintáctica de la palabra. Tras el cuarto campo, que, como ya hemos indicado, contiene la palabra exacta en el texto original, el quinto nos muestra la forma léxica base o lema de dicha palabra (*carry* es el lema de *carrying*, *see* el de *saw*, etc). En último lugar, el campo sexto es el que verdaderamente efectúa un *parsing* o etiquetado

sintáctico de la palabra en su contexto, y, en consecuencia, al que prestaremos más atención.

Limitémonos aquí a explicar las etiquetas sintácticas de las primeras palabras, así como a caracterizar la estructura general de este esquema de *parsing*. Las unidades están comprendidas entre corchetes, lo cual es bastante habitual en corpus anotados (véase Cuadro 4). Los delimitadores '[O' y 'O]' rodean lo que Sampson denomina "párrafos", que en algunos casos se corresponden más con grupos oracionales discursivamente coherentes y completos que con párrafos tipográficos. Así, *She was carrying ... from her wrist* será el primer 'párrafo' de nuestro ejemplo. Los corchetes '[Oh' y 'Oh]' se emplean como separadores de párrafos propiamente dichos. Cada párrafo está divido en oraciones, genéricamente representadas mediante '[S' y 'S]'. Hay varios tipos de oraciones: las principales, esto es, sintácticamente autónomas; las no principales introducidas por un nexo; y las no principales no introducidas por un nexo. El esquema de *SUSANNE* codifica las primeras mediante simplemente '[S'-'S]' (*She was carrying a quirt, and she started to raise it, then let it fall again and dangle from her wrist*). Las no principales encabezadas por una conjunción son aquellas comprendidas entre '[S+' y 'S+]' (*and she started to raise it, then let it fall again and dangle from her wrist*). Aquellas dependientes sin nexo inicial están encerradas mediante los corchetes '[S-' y 'S-]' (*then let it fall again and dangle from her wrist*). En cuanto al análisis pormenorizado de unidades menores, los textos del *SUSANNE* están divididos en sintagmas o frases de distinta naturaleza (nominal, verbal, preposicional, adverbial, etc.). Así, además de caracterizar *she* como una frase nominal ('[N'-'N...]'), el sistema aquí empleado nos indica que se trata de una frase nominal en función de sujeto ('Na') y en singular ('Nas'). En relación con la función sintáctica que desempeña en la cláusula, esto es, la de sujeto, el código ':s' además expresa que *she* (o el referente extralingüístico de dicha forma pronominal) actúa como el sujeto lógico (y no simplemente gramatical). Por el contrario, de la frase *a quirt* sabemos que se trata de un grupo nominal ('N') en singular ('s') en función de no-sujeto (':o'). Pasemos al grupo verbal *was carrying*, comprendido entre '[V' y 'V]'. Los códigos s y u, respectivamente, informan de que se trata de una forma verbal introducida por *was* y con estructura progresiva.

El *Penn TreeBank* de Pennsylvania tampoco es un corpus definitivo. Al contrario, se trata de un proyecto a largo plazo más centrado en el etiquetado que en el propio material textual. En la actualidad, podemos ya obtener su segundo CD-ROM, editado en 1995, que contiene aproximadamente un millón de palabras [más información en *http://www.cis.upenn.edu/~treebank*]. Reproducimos un ejemplo del procedimiento de marcaje empleado en un producto anterior, con el único objetivo de mostrar la disposición gráfica del resultado final:

```
((S
    (NP Mr. Vinken)
    (VP is
            (NP chairman
                    (PP of
                            (NP     (NP Elsevier N.V.)

                                    (NP the Dutch
                                            publishing
                                            group))))))
.)
```

Cuadro 6: *Penn TreeBank*

### 4.1.3. *Anotación prosódica*

Así como los modos de anotación categorial descritos anteriormente pueden llevarse a cabo automáticamente con mayor o menor precisión, dependiendo de las herramientas empleadas a tal fin, el etiquetado de tipo prosódico tiene que realizarse necesariamente a mano, lo que requiere gran esfuerzo y tiempo. En efecto, los corpus de lengua hablada tienen dimensiones bastante más limitadas que los de textos escritos. Transcribimos a continuación un pequeño extracto del *Corpus oral de referencia de la lengua española contemporánea* de la Universidad Autónoma de Madrid (pasaje tomado de *http://lola.lllf.uam.es/docs_es/corpus/corpus.oral/corpus.lee.html*):

```
<fuente=conversación telefónica>
<localización=Madrid> (...)
<H1=Mujer, profesora de español para extranjeros en la Universidad, (filóloga), 23 años>
<H2=Mujer, filóloga, estudiante, 23 años>
<texto>
<H1> Digo: "Bueno". Ya se me ha estropea<(d)>o el plan.
<H2> Jo.
<H1> ¡Me he podido quedar! Si me po<palabra cortada>... Me hubiese podido quedar, pero tendría que haber vuelto, ¿no? y ya me ha parecido demasiado. Digo: "Bah, ya... ya jugaré otro día con él". Porque me... al irme, me dice Alberto, dice: "¿no te quedas a jugar un poco más?"
<H2> Pero, ¿ese Alberto es... es un alumno tuyo... extranjero?
<H1> Sí.
<H2> ¿Y de dónde es, llamándose Alberto?
<H1> Italiano.
<H2> Ah, italiano. Claro, claro.
```

<H1> Di<palabra cortada>...tiene, fíjate, se llama Alberto y tiene un apellido catalán.
<H2> <risas> ¿Y eso?
<H1> ¿Eh?
<H2> ¿Y eso?
<H1> Pues no sé. Digo: "¡Pero bueno!" Dice: "Sí, sí", dice: "Si... muchas veces me he hecho pasar por español"
<H2> <risas>

Cuadro 7: *Corpus oral de referencia de la lengua española contemporánea*

En este sistema de codificación del lenguaje oral, los anotadores se han limitado a identificar el diálogo, el contexto y las personas, a asociar cada una de las intervenciones con la hablante productora y a incorporar anotaciones relativas a palabras incompletas (<palabra cortada>), reconstrucciones realizadas por el anotador (estropea≤(d)≥o) o "ruidos" técnicos de la comunicación (<risas>).

El espectro de los elementos prosódicos codificados en otros productos alcanza límites muy elevados, lo cual nos puede dar una idea del esfuerzo de sistematización y transcripción que precede a estos resultados. Un ejemplo estándar de corpus de material hablado, del que incluimos un fragmento en Cuadro 8, es el *London-Lund Corpus of Spoken English*, ya mencionado anteriormente, que comprende aproximadamente 435.000 palabras de inglés británico culto:

```
1 1 1  10  1  1 B  11 ((of ^Spanish)) . graph\ology#                    /
1 1 1  20  1  1 A  11 ^w=ell# .                                         /
1 1 1  30  1  1 A  11 ((if)) did ^y/ou _set _that# -                    /
1 1 1  40  1  1 B  11 ^well !J\oe and _I#                               /
1 1 1  50  1  1 B  11 ^set it betw\een _us#                             /
1 1 1  60  1  1 B  11 ^actually !Joe 'set the :p\aper#                  /
1 1 1  70  1  1 B  20 and *((3 to 4 sylls))*                            /
1 1 1  80  1  1 A  11 *^w=ell# .                                        /
1 1 1  90  1  1 A  11 "^m/\ay* I _ask#                                  /
1 1 1  100 1  1 A  11 ^what goes !\into that paper n/ow#                /
1 1 1  110 1  1 A  11 be^cause I !have to adv=ise# .                    /
1 1 1  120 1  1 A  21 ((a)) ^couple of people who are !d\oing [dhi: @]  /
1 1 1  130 1  1 B  11 well ^what you :dVo#                              /
1 1 1  140 1  2 B  12 ^is to - - ^this is sort of be:tween the :twVo of  /
1 1 1  140 1  1 B  12 _us#                                              /
1 1 1  150 1  1 B  11 ^what *you* :dVo#                                 /
1 1 1  160 2  1 B  23 is to ^make sure that your 'own . !c\andidate      /
1 1 1  170 1  1 A  11 *^[\m]#*                                          /
1 1 1  160 1  2(B  13 is . *.* ^that your . there`s ^something that your /
1 1 1  160 1  1(B  13 :own candidate can :hVandle# - -                  /
1 1 1  180 2  1 B  21 ((I ^won`t))                                      /
1 1 2  190 1  1 A  11 *((^y\eah#))*                                     /
```

Cuadro 8: *London-Lund Corpus of Spoken English*

Comentaremos a continuación algunas de las convenciones utilizadas en la anotación prosódica del texto anterior, dividido en líneas que contienen grupos tonales, esto es, unidades prosódicas independientes emitidas por cada uno de los hablantes (A y B) . El acento está marcado mediante los signos '^' y '!'; las marcas '\' y '/' son señales tonales; el guión '-' y la arroba '@' representan pausas en la producción (la primera absoluta, y la segunda tipo 'emm'); '#' es un límite de grupo tonal; los paréntesis '(' y ')' son indicadores de transcripciones dudosas del material sonoro; las llaves '{' y '}' delimitan unidades tonales subordinadas o dependientes, etc.

### 4.1.4. *Anotación semántica*

Hagamos uso aquí del modelo de anotación semántica empleado en la base de datos relacional de Schmidt y Wilson (más información en Wilson y Thomas 1997):

| | | | | |
|---|---|---|---|---|
| 0000001 | 001 | ——— | ——— | |
| 0000001 | 010 | NP1 | Joanna | 231112 |
| 0000001 | 020 | VVD | stubbed | 21072-31246[m1.2.1 |
| 0000001 | 030 | RP | out | 21072-31246[m1.2.2 |
| 0000001 | 040 | APPGE | her | 0 |
| 0000001 | 050 | NN1 | cigarette | 2111014 |
| 0000001 | 060 | IW | with | 0 |
| 0000001 | 070 | JJ | unnecessary | 317 |
| 0000001 | 080 | NN1 | fierceness | 227052 |
| 0000001 | 081 | . | . | |
| 0000002 | 001 | ——— | ——— | |
| 0000002 | 010 | APPGE | Her | 0 |
| 0000002 | 020 | JJ | lovely | 22706 |
| 0000002 | 030 | NN2 | eyes | 21061 |
| 0000002 | 040 | VBDR | were | 311 |
| 0000002 | 050 | JJ | defiant | 228262 |
| 0000002 | 060 | II | above | 0 |
| 0000002 | 070 | NN2 | cheeks | 2103 |
| (...) | | | | |

Cuadro 9: *Base de datos relacional de Schmidt y Wilson*

Además del marcaje de la oración (primera columna), la palabra (segunda columna) y la categoría gramatical (tercera columna), cada uno de los elementos léxicos (cuarta columna) están codificados según una determinada clasificación léxica (quinto campo). Así, '0' se emplea con palabras de bajo contenido léxico, '2103' se

refiere al cuerpo y sus partes, '21061' a la vista, '21072' a actividades físicas orientadas hacia el cuerpo, '2111014' designa objetos de lujo, etc.

El campo de la anotación semántica es posiblemente, junto con el de la discursiva (véase §4.1.5), una de las áreas de investigación menos exploradas en Lingüística de Corpus. De hecho, los proyectos de investigación relevantes en marcaje semántico, aunque de gran magnitud, son tremendamente escasos. Dejemos simplemente constancia del proyecto WordNet (Cognitive Science Laboratory de la Universidad de Princeton) y su primer producto: el etiquetado léxico-semántico parcial del *Brown* [más información en *http://cogsci.princeton.edu/~wn*].

### 4.1.5. *Anotación discursiva*

Ilustraremos el procedimiento general de la anotación discursiva mediante el siguiente pasaje tomado del proyecto de UCREL, Lancaster (más información en Garside *et al.* 1997):

| |
|---|
| (1 Feodor Baumenks 1), a former Nazi death camp guard, has asked the U.S. Supreme Court to allow <REF=1 him to retain <REF=1 his American citizenship, (2 the Hartford Courant 2) reported Monday. (2 The Newspaper 2) said (1 Federenko 1), 72, is appealing a ruling handed down (...) |

Cuadro 10: *Anotación discursiva (UCREL)*

Los números indican conexiones referenciales. Por ejemplo, *Feodor Baumenks* y *Federenko*, *the Hartford Courant* y *the Newspaper* están, respectivamente, marcados con los mismos números, esto es, '1' y '2', pues ambos pares son denominaciones de la misma realidad extralingüística. Igualmente, las proformas *him* y *his* están acompañadas por la anotación 'REF=1', que codifica el referente no pronominal. El signo '<' señala la dirección en la que dicho antecedente ha de ser hallado, esto es, una relación referente-anáfora, mientras que '>' implicaría una opuesta, es decir, catáfora-referente.

### 4.1.6. *Anotación pragmática*

La inexistencia de una gama de corpus anotados pragmáticamente nos lleva a comentar aquí el tan traído modelo de Stiles, conocido bajo la denominación de 'modo de respuesta verbal' (más información en Leech *et al.* 1997):

| | |
|---|---|
| Patient: | I have the headaches to the point where I want to<br>throw up with them (DD) |
| Dr: | Mm-hm (KK) |
| Patient: | Then I have to go to bed with them (DD)<br>And then I'll go to sleep for awhile (DE)<br>and maybe they will ease off. (ED)<br>But after a while, they'll come<br>right back again. (ED) |
| Dr: | Mm-hm (KK) |

Cuadro 11: *Anotación discursiva (modelo de Stiles)*

El modelo de Stiles se basa en la interacción entre los significados literal y pragmático. Así, 'D' codifica pensamientos, sentimientos, percepciones o intenciones; 'E' se limita a etiquetar información objetiva; 'K' es utilizado en aquellos casos de confirmación de una comunicación emitida por otro participante, etc. De este modo, el hecho de que la primera intervención del paciente esté asociada con las siglas 'DD' debe entenderse como una expresión con significado literal y pragmático de pensamiento, mientras que el empleo de, por ejemplo, 'DE' en la segunda línea de la segunda intervención del mismo personaje conduce a una interpretación literal de intención y pragmática de información objetivable, en claro contraste con secciones dialogadas anteriores y posteriores.

### 4.1.7. *Anotación estilística*

El sistema de anotación estilística *Speech & Thought Presentation*, desarrollado en Lancaster emplea etiquetas como las ejemplificadas en el siguiente cuadro (véase Leech *et al.* 1997):

```
<sptag cat=NRT>
she thinks
<sptag cat=IT>
she wins men's affections
```

Cuadro 12: *Anotación estilística (modelo Speech & Thought Presentation, Lancaster)*

La etiqueta sptag cat está asociada con valores como 'NRT' (narración de pensamiento) o 'IT' (pensamiento indirecto), ilustrado en el ejemplo anterior, u otros como 'IS' (estilo indirecto), 'FIS' (estilo indirecto libre), 'FDS' (estilo directo libre), etc.

## 4.2. *Hacia el estándar de la anotación textual: SGML-TEI*

En esta sección abordaremos el estudio de la filosofía que inspira la propuesta más universal(ista) de anotación electrónica: SGML (§4.2.1). En la segunda mitad (§4.2.2), describiremos su aplicación práctica al marcaje de textos desde una perspectiva textual, esto es, la propuesta TEI.

### 4.2.1. *SGML*

Las ventajas de poseer material textual en formato electrónico podrían reducirse a dos: la posibilidad de manejar los textos mediante ordenadores, y la versatilidad que representa el tener textos en formato interpretable por (en principio) cualquier equipo informático, y, en consecuencia, cualquier usuario y/o investigador que posea un ordenador. Si bien estas dos características por sí mismas ya justifican todo esfuerzo humano, académico y (¿por qué no?) económico destinado a la informatización de material textual, el hecho de disponer de unos principios en torno a los cuales podamos organizar esos textos ampliaría enormemente el horizonte del análisis textual computarizado. Así pues, SGML nace con la intención de convertirse en el estándar informático de la anotación.

SGML no se plantea como una alternativa a los procesadores de textos comerciales. Al contrario, SGML simplemente quiere hacerse hueco en el terreno de las "recomendaciones formales" de anotación de elementos textuales formales, las cuales muchas veces son tan inservibles en la maquetación definitiva de un texto como incapaces de conseguir los brillantes resultados de cualquiera de los paquetes de software de edición existentes en el mercado. El adjetivo "formal", utilizado más arriba como calificador de los fines de SGML, cobra un nuevo significado, pues no se refiere al aspecto externo que debe tener un elemento textual determinado, sino a la clasificación con la que dicho segmento textual se debe asociar desde un punto de vista externo. Podemos pues afirmar que SGML es una norma preocupada por la semántica del texto y no por su estética, en donde "semántica" se refiere al conjunto de conceptos formales que podemos identificar en un segmento textual (en palabras de Abaitua, "etiquetado descriptivo"), y en donde "estética" alude a la representación externa de cada uno o algunos de esos conceptos (lo que Abaitua denomina "etiquetado procedimental"). Una vez que consigamos separar la semántica y la estética en un texto, podemos asociar los resultados de una con los de la otra. Más gráficamente, podríamos disponer de una herramienta de software, más parecida a un procesador de textos o a un paquete de autoedición que a un intérprete SGML, que traduzca el marcaje SGML de una palabra o conjunto de palabras que indique "palabra extranjera" y lo exprese en cursiva, en definitiva, de una aplicación que establezca vínculos entre SGML y estándares de presentación formal de textos.

Dediquemos unas líneas ahora a otro componente más del nombre SGML: SGML como lenguaje. El objetivo de SGML no es ser un lenguaje de anotación. Es más, los

recursos gráficos que utiliza SGML para marcar un texto son perfectamente redefinibles por el propio sistema. La que aquí utilizaremos para presentar el estándar no es más que una propuesta de formalización de contenidos SGML, denominada *Reference Concrete Syntax*.(RCS), esto es, "una" sintaxis de expresión de aquello que busca reflejar el anotador de SGML. Preferiríamos, pues, caracterizar SGML como un *sistema* de marcaje de textos. Como tal sistema, incluirá un conjunto de conceptos que debemos o podemos utilizar; dispondrá de unas reglas de uso, las cuales (añadiremos) serán férreas una vez definidas; tendrá unos objetivos muy concretos; será, sin embargo, flexible según las necesidades de cada usuario. No deberá extrañarnos que el marcaje SGML de un texto puede variar muchísimo de un usuario a otro tanto cuantitativa como cualitativamente. En concreto, un usuario de SGML puede hacer uso de *n* conceptos de marcaje de un texto, frente a otro que utiliza, por ejemplo, *n*+3 conceptos. El segundo usuario puede llegar a subclasificaciones conceptuales muy estrictas y exhaustivas de cada uno de los conceptos textuales que emplea, mientras que el primero se queda en el nivel más superficial de clasificación. Ambos están haciendo uso correcto de la norma SGML. Los resultados de ambos son interpretables por un intérprete-visor-editor-analizador-gestor o *parser* de SGML, el cual los comprobará automáticamente, esto es, sin necesidad de actuación humana más que para subsanar posibles errores. En dicha comprobación, el *parser* separará el documento textual y los datos del etiquetado SGML, y se cuidará de que la anotación sea coherente y correcta.

Una vez que hemos descartado la consideración de SGML como estrictamente un lenguaje de marcaje textual, podríamos intentar definir el concepto SGML como una normativa o conjunto de normas estándar que permite a estudiosos codificar ciertos aspectos de un texto de un modo universal. El número de normas a disposición del investigador está recogido en el informe del estándar ISO 8879 de 1986. De ellas, podemos hacer uso de un conjunto reducido, o incluso ampliarlas según los intereses de cada operación de marcaje. SGML es, en definitiva, una norma estricta y a la vez flexible para los distintos fines de los investigadores. Otras normas, igualmente establecidas como estándares de proyección universal, como TEI (o su versión reducida TEI Lite), a la que nos referiremos en §4.2.2, desarrollan la filosofía SGML y la centran en disciplinas determinadas.

El estándar SGML hace uso de cuatro conceptos básicos, que ejemplificaremos más adelante mediante la normativa concreta TEI: la "entidad" de marcaje, el "elemento" de marcaje, los "atributos" del elemento de marcaje y el "tipo de documento". Hemos preferido mantener la terminología inglesa (*markup entity*, *element*, *attribute* y *document type*) en su traducción más literal a pesar de su falta de transparencia. Pasemos a reproducir las definiciones que ofrecen de estos conceptos los manuales introductorios a SGML, que serán ejemplificados mediante la propuesta TEI.

Una "entidad" de marcaje es un objeto concreto del texto: una línea, una palabra, un conjunto de caracteres, una tabla, un gráfico, un dibujo, una nota a pie de página, el

propio documento completo, etc. Estos objetos pueden estar alojados físicamente en otro ordenador, unidad de disco, CD-ROM, etc. La utilidad de este concepto reside en el hecho de que cada entidad posee una "referencia" o nombre, que sustituirá en todo momento al objeto representado. Por ejemplo, dibujo puede ser el nombre de la entidad de marcaje que se corresponde con el archivo *c:\docs\foto.jpg*.

Un "elemento" de marcaje en SGML es un concepto abstracto que sirve para denominar una realidad abstracta (denominada "objeto abstracto") en el texto. "Cualidades" como nombres, fechas, párrafos, páginas, notas a pie, etc. son "objetos" que no aparecen nombrados como tales en el texto, esto es, que no son "objetos concretos" (no son "entidades"), y que, sin embargo, pertenecen a la imagen que tenemos del texto. En SGML, mediante "elementos" representamos tales objetos pertenecientes a la realidad virtual del texto. Así, en TEI, el elemento de marcaje que delimita páginas es pb.

El concepto "atributo de un elemento de marcaje" en SGML se emplea para designar categorías de información adicional dependientes de un elemento de marcaje. Los "atributos" tienen un nombre y un valor. Por ejemplo, como ya hemos precisado anteriormente, el elemento de marcaje que se utiliza para designar el objeto abstracto 'página' en TEI recibe la etiqueta pb. El elemento pb aislado simplemente nos indicará que entre <pb> y </pb> hay una página de texto. Esta información podría completarse si añadiésemos, por ejemplo, el número de página. Esto se puede realizar mediante el atributo n, que se asocia, en este ejemplo, con un valor numérico cualquiera. Así, nos referiríamos a la página 6 mediante el marcaje:

**<pb n=6>**

**</pb>**

Uno de los conceptos de SGML más importantes y más dificultosos de codificar es el llamado "tipo de documento". En el "tipo de documento" tendremos que indicar qué entidades y elementos estamos utilizando en la anotación del texto, qué atributos están asociados con cada elemento, qué valores pueden tener los atributos de los elementos de marcaje, así como las posibles reglas de asociación entre atributos y elementos. Esta sintaxis o gramática del texto conforma el "tipo de documento". Gracias a la información ahí incorporada, el intérprete SGML puede no sólo manejar correctamente el texto, sino comprobar que la codificación interna de las entidades, elementos y atributos es correcta. En caso contrario, nos avisará del posible error, indicando la falta de coherencia entre las reglas definidas en el tipo de documento y el marcaje empleado en el texto. Además, la existencia de un tipo de documento separado y a la vez conectado electrónicamente con el documento textual facilitará la comprensión de todas las etiquetas utilizadas en el marcaje del texto por cualquier usuario o lector. Si toda la explicación y definición de entidades, elementos, atributos, valores de atributos, etc., así como de las reglas de combinación de todos esos conceptos (la "gramática" del marcaje) están codificadas electrónicamente en un

archivo especial, cualquier persona que disponga del texto, y, por tanto, de su tipo de documento, podrá seguir las convenciones del editor a la perfección. Dicho archivo es el denominado DTD o *Document Type Definition.*

Como broche final a esta sección, nos limitaremos a recomendar ciertos paquetes de distribución pública y gratuita (normalmente, disponibles a través de Internet), destinados a un público principiante en el mundo SGML, así como lugares en los que se puede encontrar material textual marcado mediante el estándar SGML:

- Información actualizada muy completa sobre aplicaciones SGML puede encontrarse en *http://www.sil.org/sgml/publicSW.html.*
- Gestores SGML (programas que permiten visionar un texto marcado según el estándar SGML, así como escribir marcaje siguiendo dicha normativa): Los gestores SGML más utilizados, sin duda, son *Panorama* de SoftQuad (versión gratuita limitada *Panorama Viewer* en *http://www.softquad.co.uk*) y Softquad Author/Editor (*http://www.softquad.com*). También gratuitos son los gestores SGML de James Clark (*sgmls*, *nsgmls*, *SP parser toolkit*, etc) y *Emacs*, disponibles desde *http://www.jclark.com/sp/index.htm* [más información sobre gestores en *http://www.sil.org/sgml/publicSW.html*]. El *BNC* posee su propio intérprete de SGML, a saber, *SARA* (véase Aston y Burnard 1997).
- Algunas bibliotecas electrónicas SGML y TEI: *Oxford Text Archive (http://firth.natcorp.ox.ac.uk/ota/public/index.shtml)*, *SGML Repository* en Oslo (*ftp://ftp.ifi.uio.no/pub/sgml*), *SGML Proyect Exeter* en Exeter (*ftp://info.ex.ac.uk/pub/sgml*), *UIC* (*ftp://ftp-tei.uic.edu/pub/tei*).
- Grupos de noticias y listas de discusión electrónica sobre SGML: *comp.text.sgml, SGML-L (listerserv@vm.urz.uni-heidelberg.de), SGML list (mailbase@mailbase.ac.uk), SGML-FAQ-L (listserv@listserv.aol.com).*
- Publicaciones electrónicas: "A Gentle Introduction to SGML" (*http://sable.ox.ac.uk/ota/teip3sg*, o *http://www-tei.uic.edu/orgs/tei/sgml/teip3sg/index.html), What is SGML and How Does It Help? (http://sable.ox.ac.uk/ota/teidw25/), The SGML Primer. SoftQuad's Quick Reference Guide to the Essentials of the Standard: The SGML Needed for Reading a DTD and Mark-up Documents and Discussing Them Reasonably (http://www.softquad.co.uk/sgmlinfo/primbody.html),* "SGML". *(http://www.deusto.es/ãbaitua/konzeptu/sgml/sgml1.htm).*

### 4.2.2. *TEI*

Como ya hemos indicado en §4.2.1, SGML ha supuesto la creación de un estándar de anotación que permite un alto grado de flexibilidad por parte del usuario. Sin embargo, el desarrollo de elementos de marcaje y atributos manejados por el editor del texto implica la realización de un DTD personal para cada ocasión. Desde esta perspectiva, la existencia de un estándar de anotación tenía que ir acompañada por la

de un estándar de DTD aplicable a la anotación de cualquier material textual. Surge así TEI (*Text Encoding Initiative*) como una propuesta universal de marcaje textual acorde con la filosofía que inspiró la propuesta de SGML.

Las normas TEI sirven para anotar información no sólo estrictamente textual, sino también sonora y visual. Todos aquellos estudiosos que se acojan al estándar TEI podrán intercambiar dicha información fácilmente. Es más, como hemos indicado anteriormente, TEI hace uso de la filosofía SGML, de ahí que cualquier herramienta de software que sea capaz de procesar textos preparados para SGML podrá también manejar textos anotados bajo TEI sin dificultad alguna. Centros de textos electrónicos como el *Oxford Text Archive* de la Universidad de Oxford (Gran Bretaña) o los centros de las Universidades norteamericanas de Virginia (*http://etext.lib.virginia.edu*) y Michigan (*http://www.hti.umich.edu*) utilizan TEI para codificar sus textos o para traducir los marcajes originales de sus archivos.

En este artículo nos limitaremos a ofrecer el comentario de un pasaje anotado siguiendo las directrices TEI. Para más información sobre TEI, un punto de partida excelente es el documento *TEI Lite: An Introduction to Text Encoding for Interchange* (documento TEI U5): disponible en *http://www-tei.uic.edu/orgs/tei/intros/teiu5.html.*

Como ha hemos señalado anteriormente, el objetivo de la propuesta TEI es el diseño de un listado amplio de conceptos que todo anotador puede, en un momento determinado, desear utilizar en el marcaje de un pasaje, texto o corpus. Por ello, hace uso de una larga lista de entidades y elementos, con sus atributos correspondientes, que cubren un amplísimo espectro de posibilidades de anotación textual. Obviamente, el ejemplo[4] que utilizaremos a continuación ilustra una mínima parte de las entidades y elementos manejados por la propuesta TEI. Sin embargo, creemos que servirá magníficamente al propósito de este artículo, esto es, dar a conocer tanto la filosofía SGML como su materialización en el caso concreto de anotación textual (TEI) desde una perspectiva sumamente práctica.

```
<TEI.x>
<text>
<front>
<tPage>
<dTitle type=main>The Last of the Mohicans</dTitle>
<dTitle type=sub>A Narrative of 1757</dTitle>
<byLine>by
<dAuthor>James Fenimore Cooper</dAuthor></byLine>
</tPage>
<div type='preface'>
<head>Introduction</head>
<pb n='vii'>
<p>It is believed that the scene of this tale, and most of the
information necessary to understand its allusions, are rendered
```

```
sufficiently obvious to the reader in the text itself, or in the
accompanying notes. Still there is so much obscurity in the Indian
traditions, and so much confusion in the Indian names, as to render
some explanation useful.
<p>Few men exhibit greater diversity, or, if we may so express it,
greater antithesis of character, than the native warrior of North
America. In war, he is daring, boastful, cunning, ruthless,
self-denying, and self-devoted; in peace, just, generous, hospitable,
revengeful, superstitious, modest, and commonly chaste. These are
qualities, it is true, which do not distinguish all alike; but they
are so far the predominating traits of these remarkable people as to
be characteristic.
(...)
</div>
</front>
<body>
<div type='chapter' n=C1> <pb n='1'>
<epigraph>
<quote><l>&odq;Mine ear is open, and my heart prepared:
<l>The worst is wordly loss thou canst unfold:&mdash;
<l>Say, is my kingdom lost&cdq;? &mdash;</quote>
<bibl rend='sc'>Shakespeare</bibl> </epigraph>
<p>It was a feature peculiar to the colonial wars of North America,
that the toils and dangers of the wilderness were to be encountered
before the adverse hosts could meet. A wide and apparently an
impervious boundary of forests severed the possessions of the hostile
provinces of France and England. The hardy colonist, and the trained
European who fought at his side, frequently expended months in
struggling against the rapids of the streams, or in effecting the
rugged passes of the mountains, in quest of an opportunity to exhibit
their courage in a more martial conflict. But, emulating the patience
and self-denial of the practiced native warriors, they learned to
overcome every difficulty; and it would seem that, in time, there was
no recess of the woods so dark, nor any secret place so lovely, that
it might claim exemption from the inroads of those who had pledged
their blood to satiate their vengeance, or to uphold the cold and
selfish policy of the distant monarchs of Europe.
(...)
</div>
</body>
</text>
</TEI.x>
```

Cuadro 13: *Ejemplo de anotación* TEI

En el formato de anotación que aquí comentaremos las secciones reales de textos, por un lado, y las entidades y elementos de marcaje, por otro, están delimitados mediante los caracteres especiales '&-;' y '<->', respectivamente. Así, el carácter '—' (guión largo), que no se encuentra en el grupo central de caracteres ASCII, será definido como la entidad *mdash*, la cual, junto con los delimitadores arriba mencionados, se convierte en el texto en — (véase ejemplos en el Cuadro anterior). Las secuencias de marcaje que codifican los conceptos en torno a los cuales son clasificadas las distintas secciones del texto (título, autor, párrafo, línea, etc.) son los llamados 'elementos', a cuyo análisis dedicaremos los siguientes párrafos.

Todo texto TEI está delimitado por el elemento <TEI>, que le sirve al analizador SGML como indicación de que el DTD que debe ser usado a la hora de interpretar la anotación del pasaje en cuestión es el correspondiente a la norma TEI. El elemento *TEI* precisa, además, la versión de la norma TEI empleada. Así, posibles formalizaciones del elemento en cuestión pudieran ser <TEI.1>, <TEI.2> o <TEI.3>.

Una anotación mediante el estándar TEI suele contar con una división superior en dos grandes conceptos, la cabecera (*teiHeader*) y el texto (*text*). La cabecera contiene información sobre la procedencia del pasaje, sobre su anotación, etc., esto es, detalles que no aparecen especificados en el texto (libro, manuscrito, artículo, etc.) que deseamos anotar. El pasaje de Cuadro 13 únicamente incluye la segunda de estas secciones, la cual abarca todo el material textual comprendido entre los delimitadores de apertura <> y los de cierre </>. Así, <text> y </text> son las marcas de inicio y final del texto en sí.

El elemento *text* suele estar formado por una sección introductoria (*front*), el núcleo del texto (*body*) y la parte final (*back*). Todas estas divisiones poseen el correlato correspondiente en TEI. La primera de ellas, el elemento *front*, suele incluir la página de títulos, los créditos editoriales, introducciones de editores, etc. *Body* delimita la parte central del documento (los capítulos de una novela, los actos de una obra de teatro, la digresión científica en un artículo académico, etc.). Finalmente, *back* está reservada para la anotación de las páginas finales, las cuales suelen consistir en detalles de imprenta, glosarios, bibliografía, etc. El pasaje de arriba posee elementos *front* (<front> </front>) y *body* (<body> </body>). En el primero han sido codificados los dos siguientes elementos: (i) la página de títulos (<tPage>), que comprende el título principal (<dTitle type=main>), el subtítulo (<dTitle type=sub>), y la línea de autoría (<byLine>) con su subelemento autor (<dAuthor>); y (ii) el prefacio ( ), con su propia cabecera (<head>). En la anotación del segundo elemento del *text*, esto es, el *body*, se ha empleado una división en capítulos ( ), el primero de los cuales incluye la cita aquí denominada *epigraph* (<epigraph> <quote>).

Detengámonos finalmente en el examen de algunos atributos de los elementos de marcaje anteriormente citados. Ya hemos aludido a la división de la página de títulos en las líneas que comprenden el título principal, el subtítulo y la autoría. En el ejemplo anterior se ha empleado un único elemento para anotar tanto el título principal como el secundario, esto es, <dTitle>. Se ha incorporado a este elemento el atributo universal TEI *type*, con la finalidad de establecer la distinción entre el principal (*main*) y el secundario (*sub*). Dicha diferenciación se consigue no mediante elementos nuevos sino mediante valores distintos del atributo *type* del elemento *dTitle*. Del mismo modo, el mismo atributo nos permite utilizar un elemento genérico como *div* para anotar cualquier división del pasaje. Así, el elemento *div* se emplea para marcar el prefacio mediante la fórmula  o los capítulos . En este último caso, un nuevo atributo universal de TEI, *n*, permite, además, incluir en la anotación del texto el cardinal correspondiente a la división. En consecuencia, el capítulo primero ha de ser anotado como una división del tipo 'capítulo' con el número '1': . El último de los atributos que aquí comentaremos clasifica la tipografía del texto que sigue al elemento del cual depende dicho atributo. Nos referimos a la fórmula de anotación <bibl rend='sc'>. El elemento destinado al marcaje de las referencias bibliográficas *bibl* está aquí acompañado por el atributo *rend*, cuyo valor *sc* (*small caps*) ofrece información sobre el hecho de que el fragmento "Shakespeare" aparece en mayúsculas inferiores o versalita. Más gráficamente, en el texto original el nombre del autor de la cita aparecería así: SHAKESPEARE. No entraremos en el comentario de los restantes elementos, dado que su significado es suficientemente transparente.

## 5. CONCLUSIONES

En este artículo hemos realizado, en primer lugar, una breve introducción histórica y conceptual a la disciplina conocida como Lingüística de Corpus. Para ello, en la primera parte de este trabajo hemos hecho un rápido recorrido por los antecedentes históricos de lo que hoy conocemos como corpus informatizados o electrónicos de textos. A continuación; sugerimos una básica taxonomía de corpus, atendiendo principalmente al tipo de información que estos incluyen. Así, las secciones principales de este trabajo estuvieron centradas en la ilustración de corpus que contienen tanto únicamente texto como, adicionalmente, anotaciones tipográficas, gramaticales, discursivas, estilísticas, etc. En la sección final, nos detuvimos en el concepto 'anotación' y en la propuesta estándar de anotación textual TEI, enmarcada en la filosofía general de marcaje SGML. Con estos elementos, el lector dispone de una visión general de los esquemas de anotación de corpus electrónicos más populares, lo cual representa el objetivo de este capítulo.

## NOTAS

1. Esta investigación ha sido financiada por el Ministerio de Educación y Cultura (MEC), a través de su Dirección General de Enseñanza Superior (DGES), proyecto PB97-0507, cuya colaboración aquí agradecemos.
2. Este artículo está basado parcialmente en la Parte I de Pérez Guerra (1998), en donde muchas de las observaciones aquí esbozadas encuentran un desarrollo más amplio y más focalizado en la aplicación de esta metodología a estudios centrados en la lengua inglesa.
3. Aunque el concepto de la representatividad de un corpus merecería atención especial, por razones de espacio nos limitaremos a sugerir una serie de referencias básicas en torno a este concepto: Biber (1990, 1993), Atkins *et al* (1992), Kretzchmar *et al* (1996) o Sánchez y Cantos (1997a, 1997b).
4. La inclusión de este fragmento anotado de *The Last of the Mohicans* cuenta con el permiso del *Oxford Text Archive* (OTA), por lo que queremos dejar aquí constancia de nuestro agradecimiento.

## REFERENCIAS BIBLIOGRÁFICAS

Abaitua, J. "SGML". [Documento de Internet disponible en http://www.deusto.es/~abaitua/ konzeptu/sgml/sgml1.htm.]

Aston, G., y L. Burnard. 1997. *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Atkins, S., J. Clear y N. Ostler. 1992. "Corpus Design Criteria". *Literary and Linguistic Computing* 7: 1-16.

Biber, D. 1990. "Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation". *Literary and Linguistic Computing* 5/4: 257-269.

Biber, D. 1993. "Representativeness in Corpus Design". *Literary and Linguistic Computing* 8: 243-57.

Biber, D., S. Conrad y R. Reppen. 1998. *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bunt, H., y M. Tomita, eds. 1996. *Recent Advances in Parsing Technology*. Dordrecht: Kluwer.

Busa, R. 1974-1980. *Index Thomisticus*. Stuttgart: Frommann-Holzboog. [en CD-ROM desde 1992].

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, Mass.: The MIT Press.

Fillmore, C. 1992. "'Corpus Linguistics' or 'Computer-Aided Armchair Linguistics'". *Directions in Corpus Linguistics*. Ed. J. Svartvik. Berlin: Mouton.

Garside, R., S. Fligelstone y S. Botley. 1997. "Discourse Annotation: Anaphoric Relations in Corpora". Eds. Garside *et al.* 66-84.

Garside, R., G. Leech y T. McEnery, eds. 1997. *Corpus Annotation. Linguistic Information from Computer Text Corpora*. London: Longman.

Garside, R., y N. Smith. 1997. "A Hybrid Grammatical Tagger CLAWS4". Eds. Garside *et al.* 102-21.

Johansson, S. 1991. "Computer Corpora in English Language Research". Eds. Johansson y Stenström. 3-13.

Johansson, S., y A-B. Stenström, eds. 1991. *English Computer Corpora. Selected Papers and Research Guide*. Berlin: Mouton.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London: Longman.

Kretzschmanr, W., Ch. Meyer y D. Ingegneri. 1997. "Uses of Inferential Statistics in Corpus Studies". Ed. Ljung. 167-78.

Lawler, J. y H. Dry. 1998. *Using Computers in Linguistics. A Practical Guide*. London: Routledge.

Leech, G., T. McEnery y M. Wynne. 1997. "Further Levels of Annotation". Eds. Garside *et al.* 85-101.

Leech, G., y R. Garside. 1991. "Running a Grammar Factory. The Production of Syntactically Analysed Corpora or 'Treebanks'". Eds. Johansson y Stenström. 15-32.

Ljung, M., ed. 1997. *Corpus-based Studies in English. Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17), Stockholm, May 15-19, 1996*. Amsterdam: Rodopi.

Marcos Marín, F. 1994. *Informática y Humanidades*. Madrid: Gredos.

Marcos Marín, F. 1996. *El comentario filológico con apoyo informático*. Madrid: Síntesis.

McEnery, T., y A. Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Pérez, Ch., A. Moreno y P. Faber (este volumen) "Lexicografía computacional y lexicografía de corpus".

Pérez Guerra, J. 1998. *Análisis computarizado de textos. Una introducción a* TACT. Vigo: Universidade de Vigo (Servicio de Publicacións).

Qiao, H.L. 1995. "The Mapping between the Parsing Annotation Schemes of the Lancaster Parsed Corpus and the Susanne Corpus". *ICAME Journal* 19: 63-91.

Sampson, G. 1995. *English for the Computer. The SUSANNE Corpus and Analytic Scheme*. Oxford: Clarendon Press.

Sánchez, A. y P. Cantos. 1997a. "Predictability of Word Forms (Types) and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus: An 8-million-word Corpus of Contemporary Spanish". *International Journal of Corpus Linguistics* 2/2: 259-80.

Sánchez, A. y P. Cantos. 1997b. "El ritmo incremental de palabras nuevas en los repertorios de textos. Estudio experimental y comparativo basado en dos corpus lingüísticos equivalentes de cuatro millones de palabras, de las lenguas inglesa y española y en cinco autores de ambas lenguas". *Atlantis* 19/1: 205-23.

Sánchez, A., *et al.* 1995. *Corpus lingüístico del español contemporáneo:* CUMBRE. Madrid: SGEL.

Souter, C., y T.F. O'Donoghue. 1991. "Probabilistic Parsing in the COMMUNAL Project". Eds. Johansson y Strenström. 33-48.

Stubbs, M. 1996. *Text and Corpus Analysis. Computer-assisted Studies of Language and Culture*. Oxford: Blackwell.

Wilson, A., y J. Thomas. 1997. "Semantic Annotation". Eds. Garside *et al.* 53-65.