# Metodología para la construcción de córpora textuales estructurados basados en XML*

**Fco. Mario Barcala**
Universidade de Vigo
Edificio Politécnico
Campus As Lagoas, s/n
32004 Ourense
barcala@uvigo.es

**Cristina Blanco**
Universidade de Santiago
de Compostela
Facultade de Filoloxía
Campus Norte
15782 Santiago de Compostela
cris@freeresearch.org

**Victor Manuel Darriba**
Universidade de Vigo
Campus As Lagoas, s/n
Edificio Politécnico
32004 Ourense
darriba@uvigo.es

**Resumen:** En este trabajo analizamos los aspectos más relevantes para definir una metodología que posibilite la construcción de córpora textuales estructurados basados en XML.
**Palabras clave:** Córpora estructurados, metodología, XML, XCES.

**Abstract:** In this article we discuss the most important issues in the definition of a methodology for the development of structured text corpora based on XML.
**Keywords:** Structured corpora, methodology, XML, XCES.

## *1. Introducción*

La construcción de córpora no es una tarea tan sencilla como en un primer momento puede parecer. Habitualmente, cuando ya se ha definido la estructura que van a tener los documentos y se han incorporado varios de ellos al corpus, surge el problema de que un nuevo documento que deseamos introducir no se amolda a la estructura que habíamos elegido. Cuando esto ocurre en una fase inicial, no es excesivamente complicado modificar y depurar esta estructura de manera que sea posible representar, sin mayores complicaciones, estos nuevos elementos que no habían sido tenidos en cuenta en un principio. Sin embargo, cuando el imprevisto aparece en etapas avanzadas, es más difícil abordar este cambio y supone un consumo de recursos difícil de asumir.

Es, por esto, por lo que cobra especial relevancia la definición de una metodología de trabajo antes de empezar a realizar cualquier tarea relacionada con el procesamiento de los textos que formarán parte del corpus. Así podrá evitarse la aparición de errores graves que supongan retrasos y elevados costes económicos y de recursos.

* Parcialmente financiado por el Ministerio de Educación y Ciencia (TIN2004-07246-C03-01), Xunta de Galicia (PGIDIT05PXIC30501PN) y Universidade de Vigo.

Existen bastantes trabajos donde se explican los procedimientos seguidos para la construcción de un corpus particular (Berber, 2004) (Sánchez et al., 1995), y otros que intentan definir algunos criterios generales para el diseño y construcción de córpora (Rodrigues, 2000) (Biber, Conrad, y Reppen, 1998) (Atkins, Clear, y Ostler, 1992), pero apenas hay trabajos donde se organicen, tanto los conceptos generales como los más concretos, para que puedan servir como orientación a cualquier equipo de trabajo en la definición de una metodología que permita la construcción de diferentes tipos de córpora.

En este trabajo analizamos los aspectos más importantes que se deben tener en cuenta a la hora de definir esta metodología de trabajo, centrándonos en los córpora textuales monolingües que utilicen el estándar XML (World Wide Web Consortium, 2004) como sistema de representación documental, siempre y cuando no incluyan ningún tipo de información lingüística adicional. Quedan al margen, por lo tanto, todas las observaciones relacionadas con la construcción de córpora anotados.

Hemos agrupado en siete epígrafes generales los factores que, en nuestra opinión, hay que considerar a la hora de definir la metodología (figura 1): objetivos del corpus, tipo de corpus, tipos de documentos, fuentes documentales, estructura de los documentos,

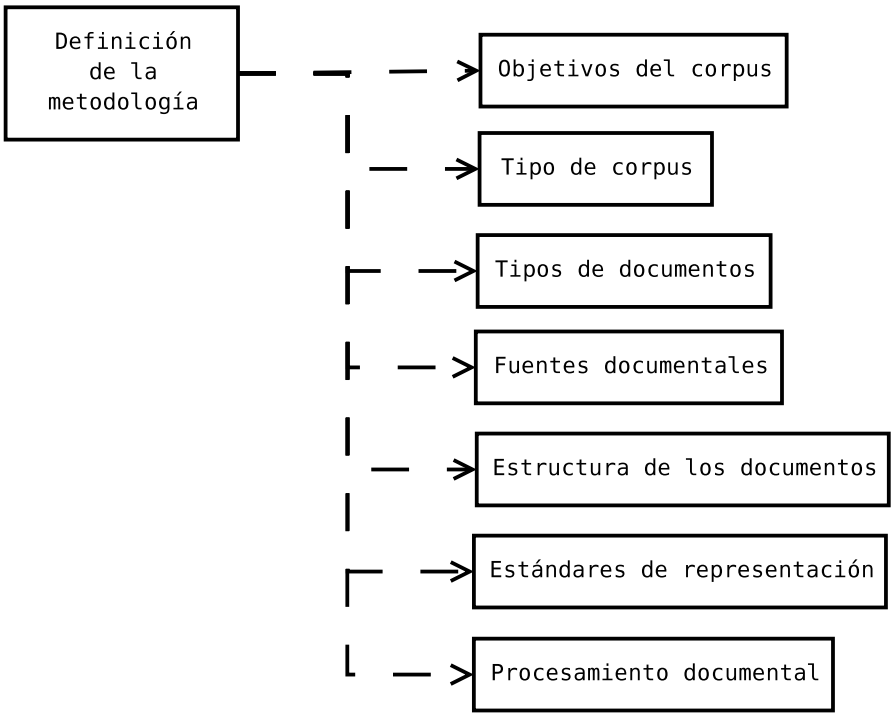


Figura 1: Aspectos clave en la definición de la metodología.

estándares de representación y procesamiento documental.

Estos factores deben analizarse y adaptarse al caso particular de utilización, por lo que los proponemos como un guión sobre el que empezar a trabajar.

## *2. Objetivos del corpus*

Los objetivos que se persiguen con el desarrollo de un corpus constituyen el primer elemento relevante a la hora de definir la metodología. Las motivaciones que animan a la construcción de córpora pueden ser de muy diversa índole, pero todas ellas convergen en el objetivo general de que posibilitan la realización de diferentes estudios, especialmente de tipo filológico (la consulta de un caso concreto de ocurrencia, la prueba o refutación de una teoría, el sondeo de casos que apoyan o rebaten una tesis, etc.).

Este objetivo general se concreta en otros más específicos, que serán determinantes cuando se definan el resto de los factores. Dependiendo del tipo de estudio que se quiera realizar, el tipo de corpus será uno u otro, la estructura de los documentos que lo compongan será de una manera o de otra, etc., por lo que es importante reflexionar concienzudamente sobre este factor para asegurarnos de que no elegimos un camino equivocado desde el principio. Por otro lado, para la identificación de los objetivos concretos, resulta de enorme utilidad pensar en el uso que se va a hacer del corpus, en las posibilidades de las herramientas existentes, o en sistemas de recuperación de información que se puedan construir *ad hoc* para el mismo.

Hay que tener especial cuidado en delimitar el alcance de nuestras aspiraciones. Inicialmente podemos imaginar multitud de objetivos particulares, pero deberemos acotarlos y priorizarlos para evitar que naufraguemos cuando se lleven a cabo todos los trabajos que dicte la metodología.

## *3. Tipo de corpus*

La segunda decisión importante consiste en determinar qué tipo de corpus vamos a desarrollar. Existen multitud de clasificaciones sobre diferentes tipos de corpus (Atkins, Clear, y Ostler, 1992) (Expert Advisory Group on Language Engineering, 1996). Basándonos en estos trabajos, y en la recopilación hecha en (Berber, 2004) y (Pérez, 2002) definimos, a continuación, los tipos de córpora textuales que consideramos más relevantes en cuanto a su influencia en la definición de la metodología:

- **Corpus de referencia**: Constituye una muestra representativa de las principales variedades de una lengua, por lo que

generalmente debe incluir una gran cantidad de documentos con el objetivo de abarcar un vocabulario general muy amplio.

- **Corpus especializado**: Corpus que se crea para que sea representativo de una variedad lingüística específica o lengua de especialidad.
- **Corpus monitor**: Originalmente (Clear, 1987) corpus de tamaño constante en el que se van incluyendo nuevos materiales al mismo tiempo que se eliminan los más antiguos. De esta forma se ofrece la posibilidad de observar cambios recientes en el uso de la lengua. Debido a los avances tecnológicos actuales, ya no es imprescindible que se vayan eliminando textos, con lo que el corpus se irá haciendo cada vez más grande.
- **Corpus de fragmentos textuales**: Corpus constituido por fragmentos de texto.
- **Corpus de textos completos**: Los textos se incluyen en su totalidad.
- **Corpus bilingüe o multilingüe**: Corpus que contiene textos de dos o más lenguas, estableciéndose algún tipo de relación entre ellos. Pueden ser textos que estén traducidos a varias lenguas (en este caso se habla de *corpus paralelo*), o textos de diferentes lenguas que compartan algunas características similares (en este caso se trata de *corpus comparable*).

El tipo de corpus que construyamos condicionará, pues, la materialización de muchas otras decisiones posteriores.

## *4. Tipos de documentos*

Antes de definir la estructura XML que tendrán los documentos es importante decidir qué tipos de documentos vamos a incorporar. Por lo tanto, habrá que tener en cuenta el género, las áreas temáticas, el medio de publicación y otros criterios específicos de relevancia para el conjunto.

Algunos parámetros que nos pueden orientar en esta decisión son los siguientes:

- **Lengua**: ¿El corpus estará constituido por textos de una sola lengua (monolingüe)? ¿Cuál? ¿Por varias (multilingüe)? ¿Qué lenguas incluirá?
- **Variedad lingüística**: ¿Necesitamos representar la variedad lingüística o, por el contrario, buscamos documentos con alguna característica que los singularice lingüísticamente?
- **Períodos**: ¿Qué franja temporal queremos abarcar con la selección de los textos?
- **Temática**: ¿Empleamos documentos de temática variada u homogénea?
- **Género**: ¿Es relevante seleccionar textos de diferente género?
- **Tipo de publicación**: ¿Nos centramos en publicaciones de un único medio o de varios?
- **Autor**: ¿Se trata de un corpus constituido por documentos de un único autor o de varios? ¿De autores que cumplen algún tipo de característica común? ¿Es relevante para el corpus quiénes sean los autores de los documentos?

Todos estos criterios, y otros más particulares que quedan fuera del análisis aquí desarrollado, afectarán y condicionarán nuestra decisión sobre qué tipos de documentos incorporaremos al corpus. Al igual que en el epígrafe anterior, examinaremos en profundidad cada uno de estos parámetros para que la decisión final sea acertada.

## *5. Fuentes documentales*

Es recomendable confeccionar una lista inicial de las fuentes de las que extraeremos los documentos. Habrá que comprobar que, efectivamente, se podrán obtener por los canales que determinemos y que respetaremos, en todo momento, la ley de propiedad intelectual cuando los procesemos y los incluyamos en el corpus.

Es imprescindible conseguir, como mínimo, uno o dos documentos de cada tipo, para utilizarlos como modelos orientativos a la hora de definir la estructura XML de los mismos.

## *6. Estándares de representación*

Actualmente, para el desarrollo de córpora utilizando XML, existen dos tendencias principales:

1. Utilizar alguno de los estándares disponibles como, por ejemplo, XCES (Department of Computer Science, Vassar

College, and Equipe Langue et Dialogue, LORIA/CNRS, 2002).

2. Definir un XML propio para la representación de los documentos.

La decisión de optar por una de estas aproximaciones depende de varios factores. Como norma general, resulta de más utilidad emplear algún estándar ya definido, ya que, al ser muchos los colectivos que lo utilizan, ofrece varias ventajas:

1. Mayor compatibilidad entre córpora. Se facilita especialmente el intercambio de información entre proyectos diferentes.
2. Se mejora la formación de los profesionales, minimizándose el impacto del coste de ésta cuando una persona se cambia de un proyecto a otro y se propicia, así, la movilidad del personal entre proyectos. Del mismo modo, se favorece la existencia de cursos de formación donde se expliquen esos estándares, ya que pueden ser impartidos por entidades ajenas a los proyectos.
3. Se optimiza el soporte de herramientas, ya que un estándar común permite desarrollar aplicaciones que pueden ser utilizadas en diferentes proyectos.

Sin embargo, existe un caso en el que un desarrollo propio de la estructura de los documentos puede resultar ventajoso: cuando consideramos beneficioso para el desarrollo del proyecto que las propias etiquetas XML estén en una lengua diferente del inglés.

En cualquier caso, y aunque nos decidamos por esta segunda opción, resulta de utilidad estudiar y analizar la estructura documental que proponen los estándares para poder, entonces, definir adecuadamente la que mejor se adapte a nuestro caso particular.

En las explicaciones posteriores supondremos que nos decantamos por la opción más compleja: definir un XML propio. Éstas observaciones serán también válidas para el primer supuesto (utilización de XCES) y las matizaremos en aquellos casos en los que no haya coincidencia entre las dos alternativas.

## *7. Estructura de los documentos*

Una vez tenemos claros los objetivos, el tipo de corpus, los tipos de documentos que vamos a incorporar, y la tecnología de representación que utilizaremos, y teniendo ya algunos ejemplos de textos candidatos a formar parte del corpus, definiremos la estructura XML que tendrán los mismos.

En un principio conviene pensar en todas las posibilidades de uso futuras del corpus, ya que es conveniente no olvidarse de ningún elemento que sea necesario en un tratamiento posterior. Sin embargo, tampoco es recomendable desarrollar, desde un principio, una estructura que cubra todas las necesidades, sobretodo las que se planifican a muy largo plazo. Normalmente, intentar abarcarlo todo desde un inicio implica que el desarrollo del corpus sea demasiado lento y que no se aprecien avances significativos durante mucho tiempo. Eso sí, para posibilitar variaciones futuras sin un coste excesivo, se tendrán en cuenta estos aspectos durante la concepción de la estructura de los documentos y se amoldará ésta de manera consecuente.

El principio básico para estructurar adecuadamente los documentos del corpus es dividir o marcar los elementos que haya que tratar, manipular o recuperar. Por ejemplo, si queremos construir un corpus que, mediante una herramienta o un sistema de recuperación de información, se pueda utilizar para buscar topónimos dentro de él, éstos deberán estar marcados de alguna manera dentro de los textos.

En la mayoría de los casos, a los diseñadores se les ocurren muchas estructuras diferentes para los documentos. Escoger la más adecuada no es fácil, pero puede servir de ayuda tener en cuenta lo siguiente:

- **La organización física del texto**: Dado que en muchas ocasiones los documentos están originalmente impresos en papel, esa forma de publicación puede influir en que no los estructuremos adecuadamente.

  Puede suceder que, en un primer análisis sobre cuál va a ser la estructura de los documentos, se llegue a la conclusión de que éstos deben estar organizados en páginas y, a su vez, las páginas en líneas. Si bien esta disposición en algunos casos está justificada porque, por ejemplo, es muy importante la maquetación de la edición, lo normal es que presente más inconvenientes que ventajas en el transcurso del desarrollo y utilización del

corpus.

Por el contrario, una organización basada en capítulos y párrafos resulta mucho más útil en la mayor parte de los casos. El conflicto aparece cuando deseamos estructurar de esta segunda forma los documentos, pero queremos incorporar igualmente la información correspondiente a las páginas y las líneas. Ya que la utilización de XML determina una estructura documental jerárquica, y que, en principio, no tenemos manera de representar varias jerarquías de un documento simultáneamente, nos vemos obligados a adoptar alguna solución de compromiso.

Existen algunas propuestas al respecto[1] que nos permiten incorporar la información de una jerarquía documental diferente sin que se vea afectada la estructura principal de los documentos, por lo que, para obtener la estructura más adecuada para los mismos, hay que considerar todas las variantes de representación, así como las diferentes técnicas disponibles para su representación.

- **El sistema de recuperación de información**: En muchos casos, la finalidad del corpus que diseñamos es la de construir un sistema de recuperación de información que permita hacer búsquedas sobre él. Como se apunta en algunos trabajos (Barcala, Molinero, y Domínguez, 2005), la estructuración de los documentos para el sistema puede diferir de la adecuada para el corpus y, en muchos casos, es recomendable que sea así.

  Por lo tanto, aunque sí es bueno tenerla presente, tampoco debemos dejar que la estructura de los documentos que tengamos pensada para el sistema de recuperación de información influya demasiado en la que consideremos para el corpus. Si fuese así tenderíamos a simplificarla y acabaría perdiendo expresividad y flexibilidad.

Dado que la organización interna de los documentos es crucial para el desarrollo acertado del corpus, debemos prestar mucha atención a no cometer ningún error grave de diseño en este punto. Si la estructura definida inicialmente es buena, no la volveremos a refinar, o sólo lo haremos en ocasiones muy puntuales y que entrañen escasas consecuencias. Ya que cuando se encuentra un error, normalmente nos encontramos en fases avanzadas del desarrollo (incluso cuando ya tenemos textos completamente procesados), a veces es bastante costoso rectificar. La magnitud de este coste dependerá, tanto de lo grave que sea el error detectado, como de lo automático que sea el proceso de corrección, que normalmente implicará un refinamiento de la estructura de los documentos.

Aunque la utilización del estándar XCES facilita un poco la tarea, también hay que decidir cuáles de las etiquetas que define son las que vamos a utilizar para los diversos tipos de documentos. XCES define etiquetas de estructuración para una variedad amplia de textos, y deberemos utilizar solamente aquéllas que necesitemos para nuestro caso concreto.

## *8. Procesamiento documental*

La metodología también incluirá y detallará las diferentes transformaciones que sufrirán los documentos, desde su formato y medio original, hasta adaptarlos a la estructura electrónica que hayamos definido para ellos. Para hacerlo, deberemos tener en cuenta:

- **El origen del documento**: Si el documento original está en papel o si lo tenemos en soporte electrónico. Dentro de los que están impresos en papel, también puede ser importante tener en cuenta la maquetación y, en los que se encuentran en soporte electrónico, el formato en el que están (*pdf*, *html*, *doc*, etc.).
- **El tipo de documento**: Dependiendo de los tipos de textos que tengamos definidos, el procesamiento de un tipo puede ser algo diferente al de otro, aunque compartan entre ellos algunas de las transformaciones.

Para cada origen y tipo de texto, es imprescindible definir y documentar un protocolo de actuación. Éste describirá detallada-

[1] Para más información, véase TEIP4, sección 31 (Text Encoding Initiative Consortium, 2002) y CES, anexo 10 (Department of Computer Science, Vassar College, and Equipe Langue et Dialogue, LORIA/CNRS, 2000).

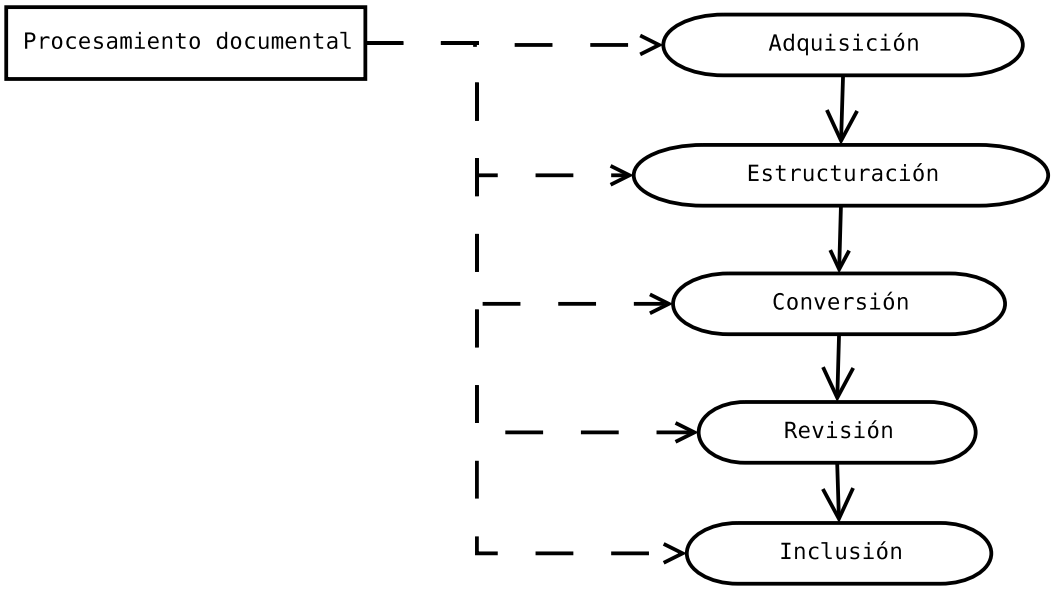


Figura 2: Fases del procesamiento de los documentos.

mente las diferentes modificaciones que se tienen que efectuar para que los textos pasen a formar parte del corpus –tanto las realizadas automáticamente por *scripts* que tendrá que desarrollar el equipo informático, como las que lleven a cabo manualmente los miembros del equipo lingüístico–. El protocolo será útil para que queden patentes todas las decisiones tomadas al respecto y para que se formen nuevos miembros del equipo de lingüistas.

Suele ocurrir que, con el paso del tiempo y dependiendo de las tareas que se estén realizando, se olviden algunos detalles de estas fases. La existencia de esta información de referencia evita volver a discutir decisiones que ya fueron tomadas en su momento y, por lo tanto, soluciona atrasos e incoherencias en el avance del proyecto.

Para realizar este procesamiento proponemos cinco fases[2] (figura 2): adquisición, estructuración, conversión, revisión e inclusión.

## 8.1. Adquisición

Se trata de una fase manual que consiste en obtener una versión digital del documento que se va a tratar. La dificultad de esta tarea puede ser muy dispar, dependiendo, principalmente, de si partimos de textos impresos en papel o en formato digital.

Incorporar un periódico digital accesible a través de la web puede ser relativamente sencillo. Llegaría con descargar las páginas que constituyen el periódico (en formato *html*, *txt*, etc.). Más sencillo aún sería que alguna editorial nos cediese una versión digital del documento que queremos incorporar (un *.doc*, un *.pdf*, etc.).

Sin embargo, incorporar un texto que solo esté disponible en papel supondría un trabajo más laborioso. En este caso sería inevitable la utilización de un programa de escaneado para ir incorporando, página a página, todo el contenido al formato digital pertinente. En este caso, el documento al finalizar la adquisición estaría constituido, bien por un fichero en el formato de la herramienta de escaneado, bien por un conjunto de imágenes con cada una de las páginas del mismo.

Por lo tanto, debemos incluir en la metodología cuál debe ser el procedimiento para realizar esta adquisición, con qué herramientas debe llevarse a cabo, etc.

## 8.2. Estructuración

Se trata de una fase manual que consiste en estructurar los documentos digitales que resultan de la adquisición, en un formato común para cada tipo de texto. Este formato común debe ser fácilmente procesable para simplificar el trabajo a los equipos informático y lingüístico, tanto en esta fase como en las siguientes. Normalmente, cuanto más sencillo es, menor será el trabajo necesario para crearlo, y

[2]Para cada corpus puede haber variaciones y/o modificaciones de alguna(s) de las fases, o incluso omisión de otra(s). En este caso, proponemos unas fases genéricas que se pueden adaptar para su utilización en diferentes tipos de corpus y sistemas de trabajo.

mayor el que se necesita para realizar las tareas de las fases posteriores. Por lo tanto, el criterio más importante a la hora de definirlo es conseguir un buen compromiso entre el trabajo que supone crearlo y el trabajo que supondrá manejarlo posteriormente.

Una posible propuesta sería utilizar una organización en carpetas con ficheros de texto. Por ejemplo, para estructurar un periódico podríamos tener un carpeta con el nombre (o una abreviatura) y la fecha del periódico. A su vez, dentro de ésta, podemos crear una subcarpeta para cada una de las secciones de noticias y, en su interior, colocar las noticias que, asimismo, contendrán un fichero con el texto de la noticia y otro con los datos sobre sus autores.

La metodología describirá, por lo tanto, cómo es el formato y cuáles son las tareas manuales que se tienen que realizar para crearlo.

### 8.3. Conversión

Consiste en una transformación automática de los documentos que resultaron de la fase de estructuración, con el objetivo de adaptarlos al formato XML que se ha definido anteriormente. Por lo tanto, se definirán las tareas que realizarán los *scripts* correspondientes para construir una primera versión de los documentos en formato XML.

### 8.4. Revisión

Dada la dificultad de la fase anterior, y teniendo en cuenta la solución de compromiso que se tiene que tomar en la fase de estructuración para determinar el formato intermedio, debemos suponer que los *scripts* no serán capaces de representar adecuadamente toda la casuística del formato XML definido para los documentos del corpus.

En la fase de revisión se completan los elementos que no han detectado los *scripts* de la fase anterior y se corrigen los errores que se hayan podido cometer. Nuevamente, cuanto más detallada sea la estructura del formato común definida en la fase de estructuración, menor será el trabajo de la de revisión y viceversa, ya que los *scripts* de la fase de conversión podrán estructurar mejor los documentos XML.

También se deben definir los protocolos de actuación que incluyan las tareas que hará el equipo de lingüistas. En ellos tienen que quedar reflejadas todas las revisiones que deben realizar las personas involucradas y el orden en el que se llevarán a cabo.

El motivo de que dividamos la creación del formato XML del documento en tres fases (estructuración, conversión y revisión) es que, de este modo, se facilitan enormemente las tareas manuales. Intentar crear este formato partiendo directamente de la versión del documento generada en la fase de adquisición, hace que se alargue considerablemente el tiempo necesario para obtenerlo.

### 8.5. Inclusión

Una vez se ha comprobado que un documento cumple todos los requisitos necesarios para formar parte del corpus, hay que hacer efectiva su inclusión, metiéndolo en la carpeta que se haya convenido. Por eso, la metodología debe definir también en qué lugar van a ser almacenados los documentos una vez procesados.

## *9. Conclusiones*

La definición de una metodología de trabajo antes de acometer el desarrollo de un corpus es de vital importancia, tanto para una adecuada marcha del proyecto a lo largo del tiempo, como para garantizar un mínimo de calidad en los documentos.

La metodología debe quedar plasmada por escrito antes de que se comience con el procesamiento de textos, de manera que se eviten improvisaciones, sorpresas, errores, olvidos y cambios de criterio durante el desarrollo del proyecto. Cuanto más detallada sea ésta, menor será la probabilidad de sufrir contratiempos y, como consecuencia, la fiabilidad del corpus mejorará.

Si bien es cierto que debemos intentar definir el mayor número de cuestiones desde un principio, es inevitable que, en un primer momento, algunas de las decisiones sean genéricas y que se hayan de ir concretando a medida que se materialicen otras. La metodología, pues, es algo dinámico, entendido el término en el sentido de que permite sucesivas concreciones y no en el de cambios de decisión sobre algún criterio, lo que como hemos visto, acarrearía graves consecuencias durante el desarrollo del corpus.

Son muchos los proyectos de construcción de córpora que han publicado los criterios y métodos seguidos durante su desarrollo, pero pocos los intentos de abstracción,

generalización y puesta en común de criterios generales que se puedan aplicar a diferentes tipos de proyecto.

Hemos intentado organizar los principales factores y procesos involucrados en la creación de diferentes tipos de córpora textuales, concretamente los que utilizan el estándar XML. Confiamos en que, en adelante, este artículo sirva como ayuda para los proyectos que trabajen en este ámbito.

### *Bibliografía*


Atkins, Sue, Jeremy Clear, y Nicholas Ostler. 1992. Corpus design criteria. En *Literary and Linguistic Computing*, volumen 7, páginas 1–16. Oxford University Press UK.

Barcala, Fco. Mario, Miguel A. Molinero, y Eva Domínguez. 2005. Information retrieval and large text corpora. En *EUROCAST 2005, Revised Selected Papers, Lecture Notes in Computer Science*, volumen 3623, páginas 91–100. Springer-Verlag.

Berber, Tony. 2004. *Lingüística de Corpus*. Manole Ltda.

Biber, Douglas, Susan Conrad, y Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Clear, J. 1987. Trawling the language: Monitor corpora. En Snell-Hornby, editor, *ZuriLEX Proceedings*, Tubingen: Franke.

Department of Computer Science, Vassar College, and Equipe Langue et Dialogue, LORIA/CNRS. 2000. Corpus Encoding Standard, version 1.5. http://www.cs.vassar.edu/CES/, 3/2006.

Department of Computer Science, Vassar College, and Equipe Langue et Dialogue, LORIA/CNRS, version 0.2. 2002. Corpus Encoding Standard for XML. http://www.cs.vassar.edu/XCES/, 3/2006.

Expert Advisory Group on Language Engineering. 1996. *Preliminary Recommendations on Corpus Typology*. EAG-TCWG-CTYP/P.

Pérez, M. Chantal. 2002. *Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento*, volumen 18. Estudios de Lingüística Española (ELiEs).

Rodrigues, José Henrique. 2000. *Introduçom à linguística com corpora (inclui estudo contrastivo cultural luso-brasileiro)*. Producción de Robinia para Edicións Laioventto.

Sánchez, Aquilino, Ramón Sarmiento, Pascual Cantos, y José Simón. 1995. *Corpus lingüístico del español contemporáneo: fundamentos, metodología y aplicaciones*. Cumbre.

Text Encoding Initiative Consortium. 2002. TEI P4: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org, 3/2006.

World Wide Web Consortium. 2004. eXtensible Markup Language (XML) 1.0. http://www.w3.org/XML, 3/2006.