# LA SISTEMATIZACIÓN TERMINOGRÁFICA: UNA PROPUESTA METODOLÓGICA PARA LA ELABORACIÓN DE DICCIONARIOS TRADUCTOLÓGICOS

Chelo Vargas Sierra
Instituto Interuniversitario de Lenguas Modernas Aplicadas, Universidad de Alicante
chelo.vargas@ua.es

## 1. Introducción

Sistematizar significar «organizar según un sistema» (RAE). Expresada la anterior definición con más concreción y aplicándola a la actividad práctica de la terminología, la sistematización implica articular, en forma ordenada y metódica, todos los elementos —fases, tareas y herramientas— que componen dicha actividad. Tal sistematización terminológica puede ser de utilidad para tres aspectos básicos: (1) dar respuesta a las cuestiones que surjan sobre el grado de efectividad de las acciones emprendidas en la elaboración de un diccionario especializado dirigido a traductores; (2) aclarar las dificultades que surgen en el desarrollo de esta actividad; y (3) favorecer la comprensión de los fundamentos implícitos y explícitos del conjunto de tareas que se llevan a cabo en el marco de una investigación. Este marco investigador se ubica en el seno de los proyectos que se están desarrollando en una de las líneas investigadoras del Instituto Interuniversitario de Lenguas Modernas Aplicadas (IULMA). Uno de sus grupos, *El Inglés Profesional y Académico,* se interesa, entre otros temas, por la investigación terminológica y la elaboración de recursos lingüísticos bilingües (diccionarios y bases de datos, principalmente) destinados al traductor de textos de especialidad. Dichos recursos sirven, desde la filosofía pragmática que aúna al equipo, para dar cuenta del uso real de las unidades léxicas de contenido especializado.

El presente trabajo pretende realizar una pequeña contribución a la mejora de la práctica terminológica bilingüe y semiautomática basada en corpus. Para ello presentaremos la organización y la sistematización del trabajo terminográfico bilingüe de carácter descriptivo y sistemático del proyecto terminológico emprendido para elaborar el *Diccionario de Términos de la Piedra Natural* e *Industrias Afines*. El principal objetivo de dicho proyecto terminológico era reorganizar la terminología de determinados ámbitos industriales de forma sistemática, moderna y dinámica, acorde con los tiempos que vivimos, que se concretase en unas herramientas lingüísticas útiles para dos tipos de profesionales: a los de la industria y a los traductores. A los primeros, en las labores de comunicación internacional y, a los segundos, para que contasen con nuevos diccionarios con los que poder optimizar su labor de mediación lingüística.

El alcance de la presente propuesta es limitado, pues se circunscribe al marco en donde se desarrolla nuestro trabajo. No obstante, pensamos que, si bien tiene sus límites, ésta puede ser de aplicación, realizando los oportunos ajustes, a proyectos terminológicos sistemáticos y descriptivos bilingües.

## 2. Las fases del trabajo terminológico

El trabajo terminológico suele seguir un proceso constante distribuido en fases. Sin embargo, estas fases no están completamente diferenciadas unas de otras, en el sentido de que no siguen una secuencia estricta. Dentro de cada fase, además, se realizan tareas destinadas a sistematizar el trabajo que se define para dicha fase. Las tareas son múltiples y muy variadas; algunas de tipo conceptual y

otras de orden documental, textual o lingüístico. En efecto, buena parte de estas tareas necesitan de una aproximación conceptual, pues es preciso entender el ámbito que se explora para poder llegar a estructurarlo, clasificarlo y definirlo. Para ello, los especialistas ayudan a los terminólogos a estructurar su área de especialidad en forma de sistema de conceptos. Esta delimitación es oportuna porque, a medida que se clasifica el conocimiento especializado, se va haciendo explícita una determinada visión cultural y científica de la realidad. El plano conceptual también está presente cuando se desea entender los términos y sus distintas relaciones con otros términos (de sinonimia, hiponimia, meronimia), así como sus significados. Además de lo anterior, es necesario recuperar textos propios de la especialidad para compilar un corpus representativo y, para ello, se han de aplicar determinados métodos que son propios de la documentación. Asimismo, al trabajar con textos, será necesario conocer los tipos textuales prototípicos del ámbito y especificar su procesamiento a nivel informático. Y, por último, el fin es capturar los términos, definirlos apropiadamente, y contextualizarlos; y aquí el terminólogo trabaja a un nivel eminentemente lingüístico, si bien no puede desvincularse totalmente del plano conceptual.

Las fases de trabajo[1] que hemos determinado son las siguientes: (1) la definición del trabajo; (2) la preparación; (3) el diseño, la construcción y la explotación de corpus; (4) la gestión terminológica; (5) la revisión y supervisión del trabajo; y (6) la edición. El gráfico siguiente representa una panorámica sintética y de conjunto de las distintas fases aludidas que configuran el flujo de trabajo:

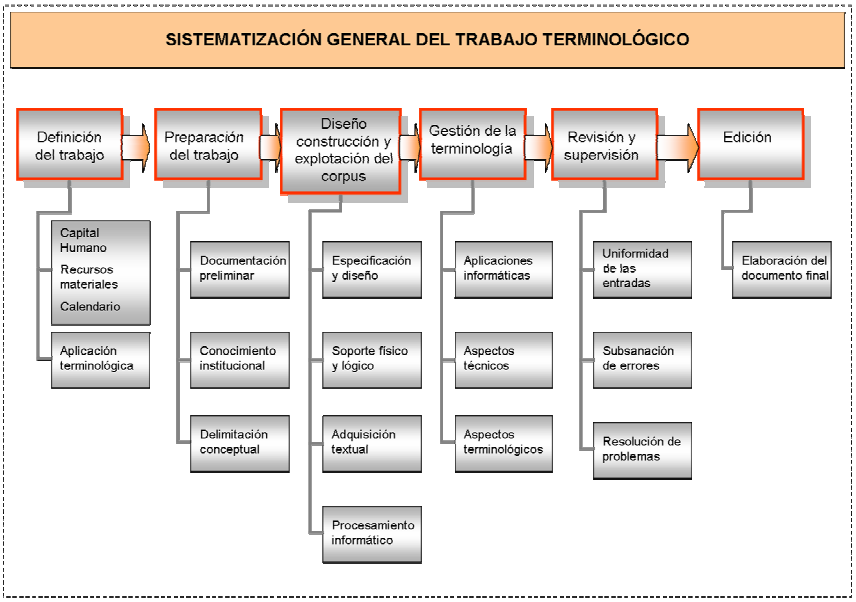


Figura 1: ***Sistematización general del trabajo terminológico (Vargas, 2005b)***

La originalidad de nuestra propuesta reside en que el diseño, la construcción y la explotación de un corpus es una fase en sí misma. Lo anterior se justifica en que la calidad de un proyecto terminológico o terminográfico está directamente relacionada con la calidad de los textos que constituyen el corpus especializado en que se basa (Bowker, 1996: 42). Si aceptamos la anterior

[1] Las fases del trabajo terminológico que configuramos como idóneas para nuestros propósitos se diseñan a partir de la asistencia de la autora a un curso en los meses de marzo a mayo de 2000, denominado «Corpus lèxics: metodología i aplicacions», organizado por el Institut Universitari de Filologia Valenciana y el Departament de Filologia Catalana de la Universidad de Alicante. Más concretamente, de una de las jornadas: «Noves tecnologies per a l'elaboración de diccionaris tècnics i científics», impartido por Lluis Rius i Alcaraz del TERMCAT, Centre de Terminología. Como es natural, las fases del trabajo que se expusieron en la aludida jornada fueron adaptadas y modificadas a nuestras necesidades.

afirmación, se requiere un modelo de trabajo terminológico en el que se establezca el contacto oportuno con las bases metodológicas constituidas por la lingüística de corpus, que nos aporta los criterios de calidad necesarios para construir una colección de textos de modo que ésta sirva para unos fines específicos de investigación. Asimismo, brinda una metodología eficaz para analizar y extraer mediante técnicas estadísticas y con herramientas informáticas los datos lingüísticos que un corpus proporciona. De este modo, los criterios de construcción de corpus y la metodología de su análisis sirven al terminógrafo para construir un corpus especializado y explotarlo de forma semiautomática. Asimismo, podemos afirmar sin miedo a equivocarnos que las etapas de construcción y de explotación de corpus pueden llegar a alcanzar más del 70% del tiempo total destinado a un proyecto terminológico. A ello hay que sumar que el material contenido en el corpus tiene una influencia directa en los resultados, como afirma Sinclair (1991: 13):

> The beginning of any corpus study is the creation of the corpus itself. The decisions that are taken about what is to be in the corpus, and how the selection is to be organized, control almost everything that happens subsequently. The results are only as good as the corpus.

Si aceptamos que la construcción y explotación del corpus terminográfico ocupa buena parte del tiempo que se destina al proyecto y que de este recurso textual dependen la calidad y los resultados de la investigación terminológica creemos que la construcción y explotación de corpus merece una fase diferenciada e independiente de otra, fase en la que se reflexione y explicite sobre cómo va a ser el corpus (tamaño, canal de producción, nivel de codificación, etc.) (*cf.* Vargas, 2006a), qué material va a contener, cómo vamos a hacer acopio de dicho material, cómo y dónde registramos los textos recogidos, cómo podemos hacer que el corpus sea representativo del ámbito objeto de estudio y equilibrado en términos pragmáticos y conceptuales, qué estrategias, técnicas e instrumentos informáticos vamos a utilizar para la extracción de datos lingüísticos, qué nivel de procesamiento informático vamos a aplicar a los textos, etc. Como se aprecia, son muchas las cuestiones que necesitan de la reflexión del terminógrafo.

Por cuestiones de espacio, en este trabajo no podemos abordar todas y cada una de las fases que constituyen nuestra propuesta. Por ello, dejaremos de lado algunos subapartados dentro de la fase de diseño, construcción y explotación de corpus, que ya fueron expuestos en detalle en un trabajo anterior (Vargas, 2006a).

### *2.1 La definición del trabajo*

La primera fase de un trabajo sistemático supone plantearse una serie de condicionantes previos o decisiones, sin cuya concreción el proyecto carecería de rigor científico. Los principales aspectos que se deben definir para el posterior desarrollo del proyecto configuran, en su conjunto, el *escenario del trabajo terminográfico*. Se trata de una fase diferente del resto, puesto que todavía no se ha comenzado con la investigación terminológica en sí, sino que se plantea una estrategia de trabajo, un plan o esquema abierto de funcionamiento interno y de operaciones que se han de realizar para crear una determinada aplicación, que, por supuesto, necesita también ser concretada. La estrategia de trabajo la concebimos de un modo abierto porque puede sufrir cambios y adaptaciones según se avance con la investigación y porque consideremos necesario optimizar cualquier aspecto que no se haya previsto, tanto en cuestiones relacionadas con las tareas, como en otras vinculadas a los recursos humanos y técnicos, el tiempo, el producto, etc.

Los elementos que componen este escenario pueden ser de diferente naturaleza, pero básicamente podemos dividirlos en dos tipos, al adoptar una óptica industrializada. Por una parte, está el producto que se pretende crear y, por otra, el capital humano (equipo de trabajo y los asesores especialistas), que a su vez se sirve de recursos materiales (económicos y técnicos o tecnológicos) para crear dicho producto y que prevé un calendario de trabajo para la consecución de sus objetivos. Los factores capital humano, recursos materiales y tiempo son considerados como variables independientes en nuestro esquema; se refuerzan mutuamente e inciden en la variable dependiente, que es la aplicación terminológica: el producto.

Los elementos iniciales que son necesarios definir con respecto al producto[2] se refieren a los factores que lo conforman y que, finalmente, harán que éste tenga unas características y no otras. En definitiva, elegiremos las que mejor se adecuen a los objetivos de nuestra investigación. Nos referimos, concretamente, a las lenguas del recurso, los usuarios o destinatarios preferentes, su finalidad, el punto de vista a partir del cual se concreta el trabajo, y su función lingüística, entre otros aspectos.

| PARÁMETROS GENERALES | SELECCIÓN DE PARÁMETROS ESPECÍFICOS | |
|---|---|---|
| lenguas | ☐ monolingüe | |
| | ☐ monolingüe con equivalencias | |
| | ☐ bilingüe | ☐ monodireccional |
| | | ☐ bidireccional |
| | ☐ bilingualizado | |
| | ☐ plurilingüe | |
| destinatario prototípico | ☐ público general | |
| | ☐ especialistas | |
| | ☐ estudiantes de la especialidad | |
| | ☐ profesores | |
| | ☐ traductores | |
| | ☐ intérpretes | |
| | ☐ lingüistas | |
| | ☐ documentalistas | |
| | ☐ redactores técnicos | |
| | ☐ correctores y revisores técnicos | |
| | ☐ otros (especificar) | |
| finalidad | ☐ consulta | |
| | ☐ comunicación entre expertos | |
| | ☐ enseñanza-aprendizaje | |

[2] Para definir una obra lexicográfica puede visitarse la página <http://terminotica.upf.es/etl/es/altres/expert.htm>. [Última fecha de consulta: agosto 2008]. Con la opción «Sistema de interrogación completo» se accede a una lista de preguntas sobre el diccionario en proyecto a las que hay que ir respondiendo. Una vez respondidas las preguntas, el programa genera automáticamente un informe en el cual se detallan las características principales del diccionario que el lexicógrafo se dispone a elaborar. Este proceso de interrogación está dividido en cinco fases: (1) Decisiones previas; (2) Superestructura; (3) Selección de informaciones; (4) Macroestructura; y (5) Microestructura. Resulta una herramienta muy útil y exhaustiva. Algunas de las consideraciones que exponemos fueron extraídas de esta página web.

| | | |
|---|---|---|
| | ☐ normalización lingüística | |
| | ☐ estandarización terminológica | |
| | ☐ documentación | |
| | ☐ divulgación | |
| | ☐ corrección/redacción de textos | |
| | ☐ traducción (directa y/o inversa) | |
| | ☐ interpretación | |
| | ☐ otras (especificar) | |
| punto de vista | ☐ especializado | |
| | ☐ semiespecializado | |
| | ☐ no especializado | |
| función lingüística | ☐ descriptiva | recoge las unidades que aparecen en los textos |
| | ☐ prescriptiva | incluye las unidades consideradas correctas o normalizadas |
| | ☐ correctiva | señala explícitamente las unidades que se han de utilizar y las unidades que se han de rechazar |

**Tabla 1:** Decisiones iniciales sobre la aplicación terminológica

Una vez definidas y concretadas las variables pragmático-lingüísticas anteriores (lengua, destinatario, finalidad, punto de vista y función lingüística), es necesario decidir tres cuestiones más (IULA, 2006):

1) Informaciones sobre la macroestructura del trabajo, desde una perspectiva general, que se pueden dividir, a su vez, en cuatro bloques:
    a. tipos de unidades terminológicas: es decir, si únicamente se van a recoger términos de distinto nivel de especialización —técnicos, semitécnicos y generales de uso frecuente—o, además, se van a registrar otros datos lingüísticos de interés para el usuario prototípico de la aplicación, como podrían ser unidades fraseológicas de núcleo preposicional (*a crédito*, *bajo fianza*), latinismos propios de la especialidad (*ab intestato*), etc.
    b. número de unidades;
    c. orden de presentación (alfabético o temático); e
    d. inclusión de anexos (bibliografía, árbol de campo, índices, etc.);
2) Informaciones sobre la microestructura o apartados de información de cada una de las entradas terminológicas, es decir, qué información lingüística, pragmática y/o conceptual se va a registrar para cada unidad. Los datos que aparecen en cada una de las entradas terminológicas se identifican según una categoría de datos, última noción ésta que se define como el resultado de la especificación de un campo de datos dado (ISO 16642: 2). La norma ISO 12620 proporciona las definiciones y los nombres normalizados para todas las posibles categorías de datos empleados en contextos de trabajo terminológico; y
3) Soporte y tipología de su edición: papel (diccionario, libro, fichero, al final de un libro, inserto en un libro de texto, etc.) o electrónico (base de datos en CD-ROM, accesible a través de Internet, documento html con hipervínculos a contextos, etc.).

## 2.2 *La preparación del trabajo*

La segunda etapa del trabajo terminográfico consiste en reunir la documentación disponible sobre el tema objeto de estudio, con el fin de acondicionar el escenario para llevar a cabo de forma más fluida el trabajo aplicado propiamente dicho. La preparación incluye las siguientes tareas: la búsqueda de documentación preliminar sobre el ámbito objeto de estudio, su conocimiento a nivel institucional y, por último, su delimitación conceptual.

Como afirma Seco (1999: XI) «para redactar un diccionario es indispensable una base documental». Todo terminógrafo que pretenda realizar un trabajo terminológico sistemático debe tener conocimientos amplios sobre aspectos como: los contenidos del tema, la documentación disponible o el medio profesional en el que el tema se desenvuelve. Roberts y Langlois (2001) apuntan que, en consonancia con la afirmación de Seco, el terminólogo se ha de apoyar siempre sobre una sólida documentación que le permita identificar los elementos que se incluirán posteriormente en las entradas terminológicas. Esta documentación es, por una parte, léxica, es decir, se ha de buscar información referente a los precedentes o diccionarios existentes de los que servirse como material básico de consulta o verificación. A esto se le conoce como *corpus de referencia* o *lexicográfico*. Por otra, la documentación que necesitamos es de orden textual, a saber: precisamos recopilar un conjunto de textos, o corpus de vaciado, que nos permita extraer los términos, estudiarlos *in vivo* y nos ayude a la hora de ilustrar sus usos, sus definiciones y otras informaciones de interés conceptual, pragmático o lingüístico. En los textos especializados se puede encontrar la mayoría de datos lingüísticos y conceptuales que sirven de materia prima al terminólogo. De la pertinencia, la diversidad, la validez y la representatividad de las fuentes dependen la calidad y la utilidad del producto final.

Además de toda la información documental sobre el tema, resulta imprescindible disponer de datos pragmáticos sobre la realidad organizativa del ámbito objeto de estudio. El conocimiento de la organización del área implicada, así como la identificación de las distintas asociaciones, instituciones, federaciones, páginas web más relevantes, etc., del campo de especialidad objeto de estudio ayudan al trabajo terminográfico en el proceso de conceptualización general y de documentación, dado que, como agentes involucrados en el ámbito que son, nos ofrecen una visión real y de calidad sobre aspectos como documentación relevante, autores de mayor prestigio y autoridad, fuentes de referencia, etc. Esta información de carácter pragmático debe almacenarse en un banco denominado 'factográfico' (Cabré, 2004).

La estructura conceptual constituye una representación de la realidad del ámbito objeto de estudio. Dicha representación pretende recoger y organizar todas las ramificaciones que son propias de un área, de modo que se refleje en forma de esquema la realidad del ámbito en cuestión. Lo anterior constituye el *sistema de conceptos* del ámbito, o sea, "el conjunto de conceptos entre los cuales o existen o se han establecido relaciones formando así un todo coherente" (Arntz y Picht, 1995: 103). Este sistema recibe también el nombre de *árbol de campo*.

No obstante, es necesario tener en cuenta la dificultad que entraña elaborar dicho sistema de conceptos, pues los ámbitos de especialidad no son construcciones cerradas, sino que evolucionan con el tiempo, presentado, por tanto, límites indefinidos, permeables y dinámicos. Como resultado, es posible elaborar más de una estructura conceptual para un ámbito determinado (*cf.* Lorente, 2001).

Con el fin de confeccionar dicha estructura, precisamos adquirir una cierta competencia cognitiva, que se puede obtener observando los contextos del corpus que se construya y estableciendo contacto y reuniones periódicas con los especialistas colaboradores. Cuando se "domina" el tema, es posible construir un primer bosquejo de lo que podría configurarse como la

estructura conceptual. Sin embargo, siempre puede haber alguien que haya realizado un primer esbozo de dicha estructura de cualquier campo. A tal efecto, los índices de manuales y los programas de asignaturas sobre el tema en cuestión pueden resultar de enorme utilidad. En el caso, por ejemplo, de que se desee trabajar en un sector de actividad económica podemos utilizar, entre otras y dependiendo de las lenguas de trabajo, dos clasificaciones: la North American Industry Classification System (NAICS)[3] y la Clasificación Nacional de Actividades Económicas (CNAE). El objetivo de ambas es categorizar las actividades productivas e institucionales de acuerdo con las actividades que desempeñan en cada país.

La Standard Industrial Classification (SIC), transformada más recientemente (1997-2002) en North American Industry Classification System (NAICS) es una de las clasificaciones jerárquicas de actividades económicas más empleadas en el mundo. Se trata de una lista numérica que incluye los principales productos y servicios de las industrias manufactureras y mineras de los Estados Unidos. Los niveles numéricos establecidos en esta jerarquía son los siguientes: dos dígitos para el sector; tres para el subsector; cuatro dígitos para el grupo industrial; cinco para la industria y seis para la industria estadounidense. La estructura de la CNAE, por su parte, viene dada por diferentes niveles de agregación, cada uno de los cuales constituye una nomenclatura. La CNAE-93[4] está estructurada de forma jerárquica en seis niveles (sección, subsección, división, grupo, clase, y subclase).

En el caso concreto del proyecto emprendido para elaborar el *Diccionario de términos de la piedra natural e industrias afines*, estas clasificaciones industriales nos resultaron muy útiles como primera aproximación para conocer los diferentes aspectos que abarcaba el ámbito que íbamos a trabajar y su organización conceptual. La exploración conceptual junto con el asesoramiento de expertos del sector nos proporcionaron los elementos necesarios para construir el árbol de campo.

Como decimos es preciso delimitar muy claramente la temática del trabajo, valorando el alcance conceptual y terminológico que la aplicación que se va a construir debe tener. La complejidad de delimitar la temática en una aplicación terminológica es aún mayor cuando se trata de materias transdisciplinares, en donde participan diversos temas transversales.

Tras una primera aproximación a la documentación relevante del ámbito, y a la realidad organizativa, temática e institucional, un modo de iniciar el trabajo de estructuración conceptual es elaborar una lista formada por una serie de descriptores temáticos. Por supuesto, esta lista no es cerrada, en el sentido de que seguramente sufrirá cambios y modificaciones dimanantes de un mejor conocimiento del área en cuestión a medida que se avanza en el trabajo. Estas marcas temáticas ayudan a trazar unas fronteras suficientemente claras, si bien en múltiples ocasiones se produce entre ellas el solapamiento; aspecto ineludible, por otra parte, cuando se pretende analizar una realidad que no puede delimitarse netamente más que por razones prácticas y metodológicas.

A continuación se presenta un ejemplo resumido de la plantilla cumplimentada en la concreción y definición de algunos de los descriptores temáticos para el sector industrial de la piedra natural.

[3] La página web donde se realizó la consulta y de donde extrajimos en su momento los datos referentes a los sectores relacionados con la industria analizada es: http://www.census.gov/epcd/naics02/naicod02.htm [Última fecha de consulta: agosto 2008].

[4] Puede obtenerse una copia de esta clasificación en <http://www.ine.es/clasifi/cnae93rev1.pdf>. [Última fecha de consulta: agosto de 2008]

| Descriptor temático | | Siglas del descriptor | | Conceptos que abarca |
|---|---|---|---|---|
| ES | EN | ES | EN | |
| Arranque | Stoping | ARRANQ | STOPING | técnicas de arranque de las rocas, arranque mecánico, de perforación y voladura, explosivos, tipos de explosivos, barrenos, perforación, escuadrado, laboreo, etc. |
| Ensayos | Tests | ENSAYO | TEST | ensayos que se realizan para la determinación de las propiedades de la piedra natural. |
| Explotación | Quarrying | EXPLOT | QUAR | diseño de canteras, tipos y partes de las canteras, métodos de explotación, de extracción de materiales, restauración de canteras y escombreras, minería subterránea. |
| Maquinaria | Machinery | MAQ | MCHN | equipos de corte con hilo, rozadoras de brazo, lanza térmica, equipos de corte con disco, equipos de chorro de agua, cuñas manuales e hidráulicas, explosivos, tecnología, maquinaria auxiliar. |
| Petrología | Petrology | PETRO | PETRO | descripción macroscópica y microscópica de las rocas, características, estructura, componentes, fisuras, alteraciones, textura, minerales constituyentes, poros, microfracturas, composición mineralógica y química, clasificación, etc. |
| Residuos | Waste | RESIDUOS | WASTE | tratamiento de lodos, aguas residuales, elementos para este tratamiento, aprovechamiento de residuos, de estériles rocosos. |

**Tabla 2:** Plantilla de descriptores temáticos

Como podemos observar en la tabla, además del descriptor temático en las dos lenguas de trabajo (español e inglés) y de los conceptos que abarca, aquí ya se decide sus respectivas abreviaturas, que son las aparecerán en la aplicación terminológica.

Pese a que la delimitación del conocimiento nunca es única y universal, las marcas temáticas resultan ser una de las informaciones más valiosas para el traductor, ya que constituyen una primera aproximación al campo de especialidad que vaya a ser objeto de traducción y un primer paso para la comprensión de las nociones que articulan el conocimiento propio de una determinada disciplina y de las relaciones conceptuales que en ella se producen.

## 2.3 Diseño, construcción y explotación del corpus terminográfico

### 2.3.1 La adquisición textual

La recuperación de textos es un proceso que requiere una gran inversión de tiempo y dinero. De tiempo, porque hay que realizar búsquedas documentales en Internet, bibliotecas, centros de documentación, bases de datos documentales, etc. Asimismo, hay textos de mucho valor terminológico (manuales, ensayos, artículos específicos) que están exclusivamente en papel y, por tanto, en primer lugar, se ha de solicitar la autorización del autor para su procesamiento; en segundo lugar, precisan un proceso de digitalización. Cuando el texto está digitalizado se revisa ortográficamente con un procesador de textos a fin de detectar los errores de reconocimiento. La inversión de dinero, viene dada, entre otros aspectos, por la necesidad de comprar el material bibliográfico seleccionado inicialmente.

La fase de adquisición textual para la posterior extracción terminológica se aproxima más a un proceso dinámico y de retroalimentación que a una fase estática e invariable. Si bien la colaboración con los expertos del ámbito resulta de gran ayuda para hacerse una idea sobre qué fuentes y publicaciones existen en la comunidad discursiva objeto de estudio y que, además, son de la confianza de sus miembros y, por tanto, idóneas para ser recogidas, la impresión del terminólogo va sufriendo cambios con respecto a las publicaciones inicialmente seleccionadas. Estos cambios en la percepción del terminólogo se producen como consecuencia de diferentes variables, a saber: la idoneidad de la información contenida, el formato o disposición que presenta el texto o la lengua en que está producido. Por ejemplo, el especialista puede recomendar las actas de un congreso determinado que el terminólogo adquiere, pero que tras una observación detenida de los contenidos éste desecha por tratar cuestiones muy específicas. Con respecto al formato, el problema surge sobre todo con los textos en papel y su proceso de digitalización. Algunas revistas pueden publicar con fondos de página que no son blancos, aspecto que dificulta en gran medida la digitalización de la muestra textual. La cuestión de la lengua del texto es un aspecto que el terminólogo también considera, dado que, *a priori*, no proporciona las mismas garantías de calidad un texto escrito en la lengua original del especialista que lo produce que otro que no cumple la anterior característica.

Si bien existen buenas razones para definir un corpus *ab initio*, se dan otras exigencias de orden práctico que pueden obligar al terminólogo a modificar algunas decisiones originales (acudir a otras fuentes, a otros tipos de texto, a otros recursos, etc. que tal vez no habían sido consideradas en un principio). En cualquier caso, creemos que el registro sistemático de la información de cada uno de los textos que se van compilando en un entorno informatizado, esto es, en una base de datos, agiliza su localización y su revisión, en caso de ser necesaria.

En la selección y clasificación de los textos que van a formar parte de un corpus que va a ser analizado lingüísticamente es necesario considerar, según la bibliografía que aborda esta cuestión, dos tipos de criterios lingüísticos: los internos y los externos. Los internos se refieren a factores puramente lingüísticos del texto y los externos tienen que ver con cuestiones extralingüísticas. En cualquier caso, tanto los aspectos internos como los externos afectan y caracterizan a los textos que nos proponemos recoger. En la selección y gestión de textos dentro de nuestro marco de investigación hemos escogido una serie de atributos textuales internos y externos, basándonos en los que se propugnan en la bibliografía que aborda esta cuestión. Sin embargo, el peso específico de nuestra selección recae sobre los parámetros externos. Lo anterior se justifica por dos razones principales, ya apuntadas por Atkins *et al.* (1992: 5):

1) los criterios externos se pueden determinar sin tener que leer el texto que nos proponemos recoger, asegurándonos, de este modo, que la selección es más objetiva; y
2) la selección inicial de los textos se ha de basar en criterios externos; únicamente después de recuperar un texto y analizarlo podríamos encontrar una serie de rasgos lingüísticos propios que contribuyen a su caracterización interna.

Los criterios externos son parámetros que se pueden obtener a partir de un análisis de la función de los textos, de sus interlocutores, la situación comunicativa en la que se producen; en definitiva, de parámetros y categorías socioculturales (Sinclair, 2003: 170) como el tema, el género, la autoría, la facticidad y la lengua.

Una vez que se ha recuperado un texto, se aplican determinados criterios de selección que nos ayudan a determinar si dicha muestra textual debe incluirse en el corpus. Los resultados de este análisis sobre cada texto se constituyen, asimismo, como los atributos básicos y elementales que son recomendables registrar para una organización y gestión óptima del corpus. Para organizar, registrar y poder recuperar los textos en razón de distintos criterios hemos diseñado un Sistema Gestor de Bases de Datos Relacional (SGBDR) en *Access* (versión *XP*) del paquete *Microsoft Office* y que denominamos *GesCorpus*[5]. Este SGBDR es fruto de un estudio previo y recoge los atributos textuales de diseño seleccionados para nuestro corpus. Se trata de una aplicación no comercializada desarrollada por la que subscribe este trabajo en el seno de nuestro proyecto.

Las bases de datos elaboradas con *Access* son, básicamente, un conjunto de tablas combinadas o relacionadas en las que se divide la información por parcelas especializadas. Una vez realizados los estudios pormenorizados pertinentes, se optó por elaborar varias tablas (un total de 12), adecuadas a la información textual que queremos registrar. Las tablas son:

1) una, de carácter principal o primaria, que es la que contiene el grueso de la información de nuestros registros, y que hemos denominado *Módulo-I*;
2) once, de carácter auxiliar o secundario, que contienen información complementaria a la principal y que son, concretamente, las siguientes:
   a) *Clasificador*: corresponde al dominio especializado de los documentos (PN, para piedra natural; CA, para el calzado; etc.). Se pueden dar de alta tantos clasificadores como sean necesarios;
   b) *Fiabilidad*: se refiere a la clasificación de la fiabilidad de un documento traducido. Los valores asignados son: alta, media, baja o ninguna;
   c) *Formato*: con los valores «papel» o «electrónico», y dentro de este último se ha de indicar la extensión del archivo (.txt, .jpg, .html, .doc, etc.);
   d) *Funciones*: función comunicativa y tenor del documento, todo ello de acuerdo con nuestra tipología textual pragmática (Vargas, 2005a). No se trata de una lista cerrada, sino que puede modificarse y ampliarse con nuevas funciones o relaciones, si fuera necesario;

[5] Se puede obtener una copia de *GesCorpus* v.2.0 solicitándola a Chelo.Vargas@ua.es.

| FUNCIONES | |
|---|---|
| **Función** | **Relación** |
| didáctico-instructivo | experto-semiexperto |
| didáctico-instructivo | experto-experto |
| divulgativo | experto-lego |
| informativo | experto-experto |
| informativo | experto-semiexperto |
| jurídico-normativo | experto-experto |
| recopilatorio | experto-experto |
| recopilatorio | experto-semiexperto |
| recopilatorio | experto-lego |

**Tabla 3:** Tabla auxiliar 'Funciones' en GesCorpus

e) *Géneros:* junto con la tabla Funciones, completa la clasificación del documento. Se pueden dar de alta tantos tipos textuales como sean necesarios;

| GÉNEROS | | |
|---|---|---|
| **Relación** | **función** | **género** |
| Experto-experto | didáctico-instructivo | instrucciones de trabajo |
| Experto-experto | didáctico-instructivo | manual de instrucciones |
| Experto-experto | didáctico-instructivo | plan de producción |
| Experto-experto | informativo | artículo de revista especializada |
| Experto-experto | informativo | informe de proyecto |
| Experto-experto | informativo | catálogo de fichas técnicas |
| Experto-experto | informativo | artículo comercial |
| Experto-experto | informativo | artículo de revista sectorial |
| Experto-experto | informativo | informe técnico |
| Experto-experto | informativo | informe de proyecto |
| Experto-experto | informativo | Monografía |
| Experto-experto | informativo | Anuario |
| Experto-experto | jurídico-normativo | Patente |
| Experto-experto | jurídico-normativo | norma de ensayo |
| Experto-experto | jurídico-normativo | normas laborales |
| Experto-experto | recopilatorio | Vocabulario |
| Experto-experto | recopilatorio | Diccionario |
| Experto-experto | recopilatorio | Glosario |
| experto-semiexperto | didáctico-instructivo | manual técnico |
| experto-semiexperto | didáctico-instructivo | libro de texto |
| experto-semiexperto | didáctico-instructivo | manual de instrucciones |
| experto-semiexperto | informativo | artículo de revista especializada |
| experto-semiexperto | informativo | artículo de revista sectorial |
| experto-semiexperto | informativo | Monografía |
| experto-semiexperto | informativo | Anuario |
| experto-semiexperto | informativo | catálogo de fichas |
| experto-semiexperto | recopilatorio | artículo enciclopédico |
| experto-semiexperto | recopilatorio | Vocabulario |
| experto-semiexperto | recopilatorio | Glosario |
| experto-semiexperto | recopilatorio | Diccionario |
| experto-lego | divulgativo | folleto publicitario |

| Relación | función | género |
|---|---|---|
| experto-lego | divulgativo | Catálogo |
| experto-lego | divulgativo | artículo divulgativo |
| experto-lego | recopilatorio | artículo enciclopédico |
| experto-lego | recopilatorio | Diccionario |
| experto-lego | recopilatorio | Glosario |
| experto-lego | recopilatorio | Vocabulario |

**Tabla 4:** Tabla auxiliar 'Géneros' en GesCorpus

f) *Lengua*: contiene un catálogo de idiomas que complementan tanto a los documentos principales como a sus paralelos;
g) *Muestras*: compuesta de siete campos: documento[6] (nombre del archivo) muestra (número de muestra), palabras (número de palabras), página inicial, página final, tema y subcampo;
h) *Paralelos*: contiene tres campos: un identificador relacionado con el documento original, lengua del documento paralelo y el nombre del archivo asignado al documento paralelo;
i) *Tenor*: relación entre productor-receptor del texto (experto→experto; experto→semiexperto; y experto→lego);
j) *Procesador*. nombre del investigador que ha procesado el documento en todas sus fases;
k) *Subcampos*: relación de todos los subcampos, fruto de la elaboración del árbol de campo.

El objetivo era confeccionar una herramienta informática fácil de usar o intuitiva para los usuarios. Por ello, se elaboró un formulario que, en cierto modo, enmascara la "parte dura" de un SGBDR hasta convertirlo en una serie de pantallas ante las que el procesador de una muestra interactúa de una manera sencilla y práctica.

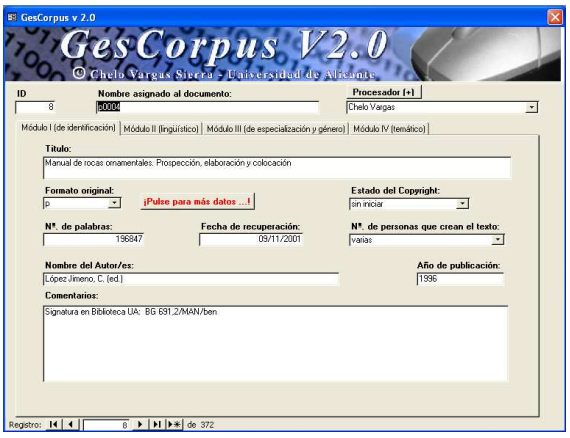


**Figura 2:** Pantalla principal de GesCorpus

[6] El documento es el texto original, con sus gráficos, ilustraciones, índice, etc. Una muestra es un archivo electrónico que se corresponde con un documento entero, o bien con una de sus partes. Por ejemplo, un artículo especializado se convierte, por lo general, en una única muestra electrónica. Sin embargo, Los documentos grandes (manuales, monografías, etc.) se fragmentan en varias muestras. Así, para un libro con varios capítulos se realiza una muestra para cada capítulo.

En la imagen anterior observamos que hay varios apartados. En la parte superior, justo debajo de la imagen, aparece información de carácter general del registro y contiene tres campos: (1) el 'ID' (identificador que asigna automáticamente la base de datos a un registro); (2) el 'Nombre asignado al documento' (el ejemplo es 'p0004'); y (3) el 'Procesador' (nombre del terminólogo que procesa la muestra). En esta misma pantalla, se aprecia que la información está dividida en cuatro grandes bloques de información, que se corresponden con el título de cada una de las pestañas de la parte superior. Los cuatro módulos son:

- *Módulo I (de identificación)*: en este apartado se registra aquella información que identifica el documento (Figura 2). Se compone a su vez de los siguientes campos:
    - Título del documento original;
    - Formato: aquí se indica si el documento estaba originalmente en papel o en formato electrónico;
    - Estado del *copyright*: hemos previsto cuatro valores: sin iniciar, iniciado, rechazado y resuelto;
    - Número de palabras del documento;
    - Fecha de recuperación del documento;
    - Número de personas que crean el texto, con tres valores; una, varias y desconocido;
    - Nombre del autor o autores;
    - Año de publicación;
    - Comentarios;
- *Módulo II (lingüístico)*: en este apartado se refleja la información relativa a los aspectos lingüísticos del documento:

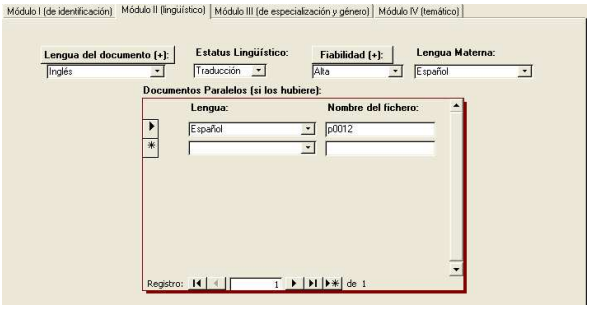


**Figura 3:** Módulo Lingüístico en GesCorpus

El módulo lingüístico se compone, como podemos apreciar en la figura anterior, de varios campos:

- Lengua del documento;
- Estatus lingüístico: con dos valores: original o traducción;
- Fiabilidad: en donde indicamos el grado de fiabilidad que nos merece un documento traducido (alta, baja, media o sin determinar);
- Lengua materna del autor del documento: si no se puede llegar a determinar se indica entonces «desconocido»;
- Documentos paralelos: consiste en un subformulario que admite tantos valores como deseemos y en el que debemos reflejar los datos correspondientes a la lengua del documento paralelo y el nombre asignado al fichero paralelo. De este modo establecemos un vínculo lingüístico entre el texto original y su traducción;

- *Módulo III (de especialización y género):* en este apartado registramos la información relativa al tenor, a las funciones del texto y al género al que pertenece, todo ello de acuerdo con una tipología textual pragmática diseñada (*cf.* Vargas, 2005a);

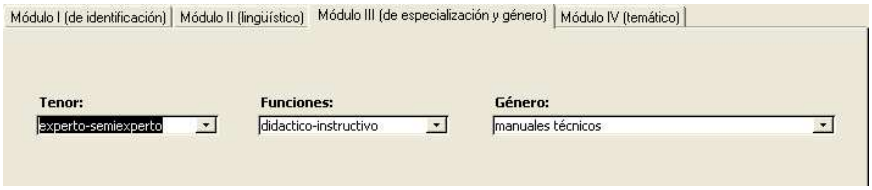


**Figura 4:** Módulo de especialización y género en Gescorpus

- *Módulo IV (temático)*: en este módulo indicamos el tema sobre el que versa el documento, el número de muestras del que se compone, así como su subdominio, de acuerdo con el árbol de campo configurado, como se observa en la siguiente figura:

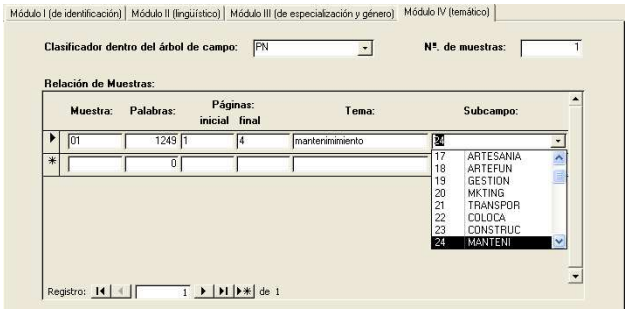


**Figura 5:** Módulo temático en GesCorpus

El tener registrados en una base de datos los atributos de cada uno de los textos permite, a través de la opción de 'Consultas' de *Access*, recuperar información ordenada de algún aspecto que nos interese sobre nuestro corpus. De este modo, podemos ir observando cómo evoluciona, si los subdominios se van representando de forma equilibrada, así como los distintos niveles de especialización, con cuántos textos paralelos contamos, cuántos en cada idioma, el número de palabras de cada lengua, etc. Podemos, del mismo modo, crear subcorpus por subdominios, por tipo de texto, por funciones, por tenor, por lenguas, entre otras opciones. En definitiva, el registro informatizado de los atributos textuales definidos por el terminógrafo para su corpus no sólo facilita la reutilización del corpus a través de la elaboración de subcorpus, sino que también resulta un medio sistemático de obtener resultados, que son, a su vez, más fácilmente interpretables.

*GesCorpus* es una base de datos adaptada a los propósitos específicos de nuestros proyectos, esto es, la compilación de corpus especializados bilingües con fines terminográficos. Su diseño, por tanto, se configura para responder a nuestras necesidades y a los atributos textuales definidos. Ello no implica que no pueda ser adaptada a otros proyectos de corpus; más bien lo contrario, pues se trata de un sistema gratuito, abierto y flexible que puede ser útil si se realizan las modificaciones pertinentes en cada caso.

Es importante también hacer constar que en el momento que se tiene físicamente la muestra textual que va a ser procesada se inicia la cumplimentación manual de un formulario que denominamos *Control de documentos*. Se trata de la versión en papel de *GesCorpus* y que da soporte a la introducción posterior en nuestra base de datos de los campos completados en la versión papel. Justificamos una cumplimentación manual en papel de los datos de la muestra textual que se va a procesar por cuestiones de eficacia. Una muestra puede no ser procesada en un único día; lo anterior depende del tamaño del documento, del número de muestras seleccionadas de éste, de su

formato original, etc. Resulta más sencillo ir recopilando los datos del documento de forma manual y una vez que se termina el procesamiento de la muestra en su integridad se registra sus detalles en *GesCorpus*. Asimismo, se guarda una copia impresa de las muestras escogidas del documento junto con las páginas que contienen los datos de identificación del documento y el índice, si lo hay. Estas copias se archivan junto con su correspondiente control de documentos (Anexo I):

### 2.3.2 *Procesamiento informático de los textos: adquisición y registro*

La lengua que emplean los diferentes miembros de una comunidad científica queda patente a través de la expresión oral y escrita. Estas formas escritas o transcritas, si eran originalmente orales, son las que se guardan y almacenan en lo que se conoce como *ficheros textuales*. Pero una mera colección masiva de ficheros lingüísticos no resulta, por sí misma, de gran ayuda al investigador; es fundamental analizarla y explotarla con herramientas informáticas.

Una vez que el texto es recuperado y seleccionado según los criterios definidos y relevantes de nuestros proyectos de corpus, la muestra textual que se va a incorporar necesita, como paso previo, ser adecuada en cuanto a su formato para que pueda ser interpretada por los programas informáticos que utilizamos, esto es, debe ser compatible con dichos programas.

El texto original puede hallarse bien en papel, o bien en formato electrónico. Obviamente, las tareas que se han de realizar en cada caso son diferentes. En el siguiente gráfico se puede apreciar de forma esquemática las fases de adecuación y registro por las que pasan los textos en nuestro proyecto para ser incorporados adecuadamente al corpus y antes de proceder con la extracción de terminología:

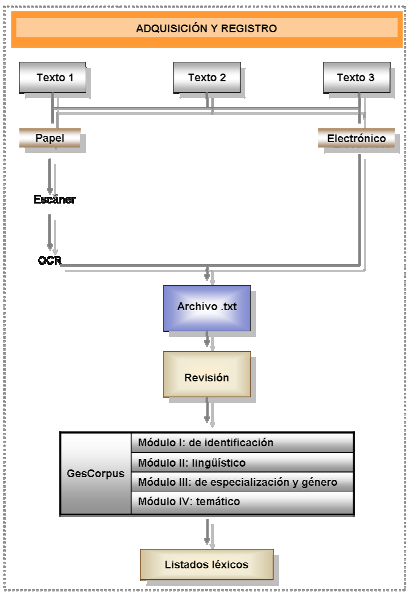


**Figura 6:** Proceso de adquisición y registro textual

### 2.3.3 *Procesamiento informático de los textos: análisis textual*

El análisis textual que se realiza en nuestro marco investigador se fundamenta en tres pilares básicos: un método, unas técnicas y unas herramientas.

Con respecto al *método*, se hace necesario aludir a que la investigación lingüística que utiliza un corpus como objeto de estudio trabaja dentro de un marco que comporta dos vertientes básicas: el enfoque basado en corpus (*corpus-based*) (Aarts, 1991; Leech, 1991) y el enfoque dirigido hacia el corpus (*corpus-driven*) (Sinclair, 1991). A pesar de la existencia de estudios establecidos en algunos de los dos polos teóricos, estas dos posturas no tienen un carácter excluyente y, de hecho, se realizan estudios utilizando ambos enfoques en la investigación.

El enfoque basado en corpus es de naturaleza confirmatoria, en donde la evidencia de los datos puede asumir un papel de *ejemplo* que apoya o invalida una teoría anteriormente formulada por el investigador. Independientemente de los resultados del estudio, el investigador no cuestiona las unidades y categorías preexistentes tradicionales. Los principales usuarios del corpus que adoptan este enfoque investigador son, por lo general, lingüistas aplicados que se interesan en desarrollar nuevos métodos para la enseñanza de lenguas extranjeras (Pearson, 1998: 50).

El enfoque dirigido hacia el corpus, por su parte, es de tipo exploratorio y empírico, en el cual el lingüista busca en el corpus patrones y distinciones entre los datos en los que basar la formulación de generalizaciones lingüísticas y, de este modo, llegar a una teoría. Desde aquí, el investigador no tiene ideas preconcebidas sobre la validez de una teoría específica para la descripción lingüística; cualquier conclusión se realiza sobre la base de las observaciones directas a partir de la lengua en contexto, esto es, en virtud de la presencia o ausencia en el corpus de construcciones lingüísticas recurrentes, su uso y su frecuencia. Este último enfoque es el que adoptan usuarios del corpus como lexicógrafos, terminógrafos y lingüistas computacionales, pues uno de sus objetivos es descubrir nuevos datos sobre una lengua o variedad de la misma (*ibid*: 49).

En lo referente a las *técnicas*, recurrimos a dos, similares a las propuestas por Ahmad *et al*. (1994: 268). La primera parte del supuesto de que la lengua tiene una naturaleza probabilística (Halliday, 1991: 31), por lo que empleamos diferentes métodos estadísticos para identificar posibles términos simples y combinaciones terminológicas. La segunda se basa en la explotación de claves lingüísticas en el texto para identificar términos potenciales y términos relacionados semánticamente (sinónimos, hipónimos, merónimos, etc.). Clear (1993: 273) describe de forma muy reveladora y sintética en qué consiste la aplicación de los métodos estadísticos, tanto en lo referente a las formas simples como al hallazgo de lo que constituye una colocación:

> The data-driven approach to collocation proceeds from the study of word-forms, their frequency, the frequency of their co-occurrence with the other forms and the statistical measurement of the significance of such co-occurrence.

Partiendo de la anterior descripción de la técnica estadística, y aplicándola a la identificación de términos, asumimos que estas unidades aparecen con una frecuencia elevada en los corpus especializados que construimos, aspecto que nos permite identificar los términos simples. Con respecto a las combinaciones terminológicas, partimos de que si una unidad léxica *X* aparece de forma frecuente en el entorno o distancia (*span*) de una unidad léxica *Y* que de otra manera en el corpus, entonces es que *X* e *Y* forman, *a primera vista*, una combinación significativa. Este es el principio denominado *información mutua* (IM). Subrayamos *a primera vista* porque somos conscientes, no obstante, de que se necesita realizar otro tipo de pruebas a fin de cerciorarnos de que, efectivamente, nos hallamos ante una unidad o combinación susceptible de seguir siendo procesada. Es claro que el análisis estadístico por sí mismo es insuficiente para etiquetar un término como tal o para identificar una combinación terminológica, pero sí sirve como técnica previa que puede ayudar a su identificación inicial y extracción semiautomática.

Otro apunte necesario tiene que ver con las *herramientas*. Como es sabido, la utilización del corpus está en íntima relación con los diferentes procesos para su tratamiento y con las

herramientas utilizadas en dichos procesos. De este modo es como se hace posible su uso y su explotación. Los instrumentos informáticos para el tratamiento y análisis de corpus se pueden dividir en dos grandes grupos: los etiquetadores morfosintácticos o desambiguadores, y los programas de análisis textual, más popularmente conocidos por *programas de concordancias*[7]. Los primeros llevan a cabo automáticamente la identificación y asignación de las categorías gramaticales de todas las palabras de un texto. Los segundos, por su parte, son un conjunto de aplicaciones informáticas capaces de analizar una base de datos textual (el corpus) y generar diferentes tipos de listas de los resultados (*cf.* Vargas, 2006b). En definitiva, facilitan su explotación, puesto que ofrecen diversos modos de visualizar y gestionar los resultados producidos por una consulta específica.

El tratamiento informatizado de un corpus no etiquetado es mucho más limitado que uno etiquetado, pues al no contener todas las unidades del corpus marcas descriptivas es imposible realizar búsquedas por patrones lingüísticos. A cambio, puede ser explotado por herramientas más generales. Aquí se incluyen los programas de concordancias, que llevan a cabo tareas como generar listados de palabras y estadísticos, producir líneas de concordancias, de agrupaciones léxicas, entre otras.

Uno de los factores que puede favorecer un mayor uso de herramientas informáticas en el análisis lingüístico es la existencia de programas flexibles y fáciles de utilizar que, además, estén bien comercializados, o bien sean de libre distribución. En este sentido, uno de los programas que cumple estas exigencias es *WordSmith Tools*, paquete informático desarrollado por M. Scott (1997) de la Universidad de Liverpool y distribuido por Oxford University Press. Cuenta ya con más de nueve años de existencia y su versión actual es la 5.0[8].

### 2.3.2 Gestión de la terminología

La manipulación, almacenamiento y recuperación de los datos terminológicos se lleva a cabo utilizando los sistemas informáticos de gestión terminológica, también denominados *sistemas gestores de bases de datos terminológicas* (SGBDT), o, de forma sintética, *bases de datos terminológicas* (BDT). Por tanto, los podemos definir como herramientas o paquetes informáticos que están diseñados específicamente para gestionar datos terminológicos. *Gestionar* implica que el usuario puede llevar a cabo las tareas de recopilación, almacenamiento, manipulación y recuperación de los datos terminológicos. En un sentido amplio, una BDT es un sistema informatizado de almacenamiento y gestión de unidades léxicas de contenido especializado que se estructuran de acuerdo con determinados criterios, con los usuarios y con la finalidad de la compilación terminológica. Para llevar a cabo dichas acciones, la BDT debe cumplir con cinco premisas básicas: (1) ser flexible; (2) reflejar adecuadamente las relaciones entre jerarquías de información; (3) permitir el almacenamiento de todos los datos pertinentes; (4) poder recuperar fácil y rápidamente la información; (5) ofrecer distintas posibilidades de presentación.

Dentro de esta fase, son dos las decisiones generales que se han de tomar: qué SGBDT vamos a utilizar para gestionar las fichas o registros y qué campos o apartados de información va a contener cada una, que deben, por supuesto, adecuarse a los parámetros originales diseñados para la

[7] En la actualidad los términos *programa de análisis textual* y *programa de concordancias* se utilizan de manera indistinta. Podríamos decir que el primero es el hiperónimo, pues estos programas también incluyen una utilidad para realizar concordancias. Sin embargo, los programas informáticos que se conocen por *programas de concordancias* también llevan a cabo otro tipo de análisis sobre los textos, que pueden estar o no etiquetados.

[8] En Vargas (2006b) puede encontrarse una descripción detallada del programa y el modo en que lo utilizamos en el análisis lingüístico de los corpus.

microestructura de la entrada (informaciones lingüísticas, pragmáticas y conceptuales) en la fase de definición del trabajo.

Antes de decidirse por un SGBDT concreto es necesario evaluar las diferentes herramientas existentes siguiendo criterios exhaustivos y adecuados a las necesidades concretas de un proyecto dado. De este modo, será posible obtener una visión real de su aplicabilidad al contexto específico de trabajo. Un SGBDT puede evaluarse siguiendo múltiples criterios, todos ellos de distinto nivel de importancia, pero que merecen, en cualquier caso, cierto tiempo de consideración, puesto que si la herramienta seleccionada no se adecua a las necesidades concretas y específicas del proyecto se pierde tiempo, dinero y esfuerzos. Por cuestiones de espacio no es posible especificar cada una de las cuestiones que se deben plantear y responder antes de decidirse por un sistema concreto. No obstante, a modo de resumen, se pueden destacar las que se relacionan a continuación (EAGLES, 1995; GTW, 1996):

1) Descripción técnica: requisitos de hardware y software para que la herramienta funcione perfectamente.
2) Compatibilidad con versiones anteriores, y con las posibles actualizaciones.
3) Interfaz de usuario: cuál es el procedimiento de instalación, tipo de interfaz, lenguas en las que se visualiza, tipo de ayuda a disposición del usuario y modo de acceder a ella (manual, en línea, soporte técnico telefónico/e-mail/en línea, a través de listas de distribución o foros, por ejemplo), en qué lenguas se presenta la ayuda, modo de representación de la información en pantalla, posibilidades de manipulación de los menús, iconos, botones, de la visualización de la información en pantalla etc.
4) Aspectos terminológicos: esta es una de las cuestiones más importantes y prácticamente la que rige el resto. Se divide, a su vez, en dos aspectos fundamentales: (a) la gestión de la terminología (lenguas que admite, modelo de datos de la base de datos [relacional, orientado a objetos, de red semántica], tipo de datos que se pueden introducir [textuales, gráficos, multimedia], número máximo de bases de datos, posibilidad de cambiar la dirección lingüística, de abrir varias bases de datos al mismo tiempo, etc.); y (b) el modelo y estructura de la entrada: evaluar si cumple los estándares, si se trata de una estructura de ficha fija, libre o semilibre, si es posible añadir o modificar campos, cambiarles el nombre, si admite categorías de datos administrativos, si hay distintos tipos de campos, el número de campos máximo y longitud de éstos en número de caracteres, posibilidad y modo de crear referencias cruzadas entre entradas, y número total posible de registros para cada base de datos.
5) Extracción de información: modo de consulta, modo en que el sistema responde a las consultas, posibilidad de restringir el acceso a la base de datos.
6) Introducción de los datos: cuestiones relativas a la edición (posibilidad de dar formato a los caracteres, de copiar y pegar cuando se edita la entrada, de configurar distintos diseños de edición ...), validación/control (revisión ortográfica, aviso de entradas duplicadas o de la omisión de un campo obligatorio, etc.).
7) Intercambio de los datos terminológicos: impresión (posibilidad de impresión de los datos en su conjunto o de una selección, de configurar el diseño de impresión), proceso de importación/exportación (posibilidad de definir los criterios de importación y/o exportación, de configurar las vistas, formatos de los ficheros de importación que admite, realiza estos procesos con formatos estándar (TMX, Martif).
8) Interacción con otras aplicaciones: con procesadores de textos, con otras bases de datos y aplicaciones, etc.

9) Fuentes (tipos, estilos) y mapa de caracteres disponible.
10) Operaciones necesarias para su mantenimiento.
11) Aspectos comerciales; fabricante, distribuidor, precio, disponibilidad y servicio técnico en el territorio nacional, personas/instituciones conocidas que lo utilicen, entre otros.

Con respecto al registro terminológico, éste debe comprender una gama de categorías de datos, todas independientes, que sigan los principios de la norma ISO 12620:1999, al tiempo que se adecuen a las informaciones que se quieren incluir sobre los términos. En nuestro caso, por tratarse de proyectos bilingües (inglés-español), los registros que configuramos están formados por dos módulos lingüísticos, cada uno de ellos, a su vez, compuesto por el mismo número de campos. La configuración del diseño de nuestras fichas electrónicas se plasma a partir de la siguiente estructura básica:

1) **Datos administrativos:** contiene campos compartidos para todas las lenguas del registro. Son registrados o asignados de forma automática por el sistema gestor elegido. De entre los posibles, seleccionamos:
    - Número de ficha o registro.
    - Proyecto (se indica el nombre).
    - Diccionario (se consigna el nombre).
    - Creado por.
    - Modificado por.
    - Creado el.
    - Modificado el.
    - Área temática.
2) **Datos terminológicos**; en esta categoría se incluyen:
    - datos lingüísticos: el término o la entrada, la categoría gramatical, las abreviaturas y las colocaciones;
    - datos pragmáticos: marcas de uso, marcas de normalización, marcas de variación lingüística y el contexto (ilustrador del uso del término).
3) **Datos bibliográficos:** en este campo se consigna la fuente de donde se extraen distintas informaciones. Son tres los campos que creamos para incluir información sobre las fuentes:
    - fuente del dato (para registrar de dónde se ha extraído el término);
    - fuente de la definición (para registrar la procedencia o la autoría de la definición); y
    - fuente del contexto (para registrar la procedencia del contexto).
4) **Datos conceptuales:**
    - la definición;
    - el descriptor temático o subdominio al que pertenece el término dentro de la estructura conceptual creada;
    - las referencias cruzadas o conceptos relacionados; y
    - los sinónimos.
    - gráfico/imagen.

En el SGBDT elegido (*TermStar*) la estructura de la ficha es jerárquica y se divide en dos partes: el encabezado y la entrada. Los datos que se almacenan en el encabezado son los de tipo administrativo (nombre del proyecto, número de concepto, fecha de creación...) y los comparten todas las lenguas de trabajo que se configuren y formen parte de la base de datos. En el encabezado también es donde se incluyen los gráficos. Por otra parte, los datos que se almacenan en la entrada

son de tipo terminológico, bibliográfico o conceptual, y el número de lenguas de trabajo es ilimitado.

A partir de las anteriores categorías de datos seleccionadas y de las posibilidades de diseño de un registro que ofrece el SGBDT elegido, la ficha terminológica modelo de nuestro equipo de trabajo se materializa de la siguiente manera:

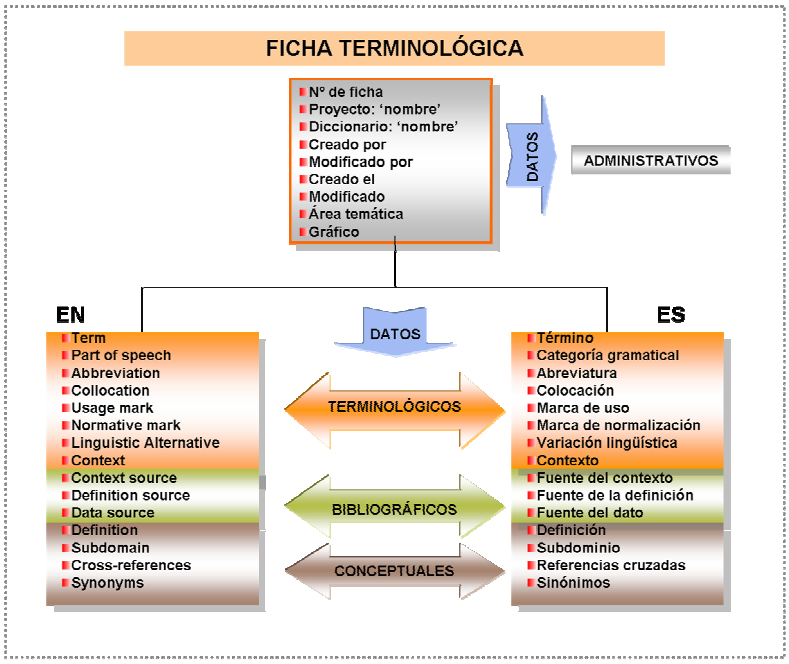


**Figura 7:** Ficha terminológica modelo

Una vez seleccionado el SGBDT y los campos que contendrá la ficha, el siguiente paso consiste en profundizar sobre ellos y definir cuestiones de funcionamiento, de cumplimentación y de contenido en su nivel interno. En la definición de la estructura interna de cada campo de las fichas, en nuestro caso concreto, hemos decidido incluir una serie de elementos, unos optativos y otros obligatorios, teniendo en cuenta la finalidad del producto, las posibilidades de la herramienta seleccionada y el destinatario prototípico: el traductor.

Cabría añadir que para incluir la información terminológica en la ficha de forma homogénea y sistemática entre todos los que componemos el equipo de trabajo es recomendable elaborar una guía o *protocolo de redacción.* En nuestro caso, este documento sigue los principios propuestos por Pavel y Nogel (2001) y TERMCAT (1990), principios que hemos adaptamos a nuestros objetivos y circunstancias concretas. En ella se describe, por ejemplo, qué tipo de contextos nos interesa incluir, el modo de representación de las fuentes bibliográficas, el estilo de redacción de las definiciones, las fórmulas preferidas para redactar las definiciones en determinadas categorías gramaticales (conceptos de estado, de acción, adjetivales y locuciones), se especifican las marcas de uso, de normalización, de variación lingüística, de frecuencia, temporal, de ponderación, geográficas, etc.

### 2.3.4 *Revisión y supervisión*

Esta fase incluye una revisión y supervisión final de la aplicación terminológica por parte del terminógrafo y de los expertos colaboradores del ámbito explorado. Estos últimos intervienen en

cuestiones más bien conceptuales, pues advierten con relativa facilidad cualquier error conceptual, la ausencia de términos que se consideran importantes del ámbito, o bien pueden sugerir la eliminación de algún término por considerarlo no pertinente o innecesario. Los aspectos lingüísticos se reservan al terminógrafo, pues su formación en este campo les permite realizar valoraciones más precisas sobre la forma en que están redactadas las definiciones, la sintaxis apropiada de los contextos o la validez de los equivalentes. El objetivo primordial de esta fase es asegurar que los datos terminológicos sean precisos y completos, al tiempo que satisfacen las necesidades del usuario final. También hay que intentar dar respuesta a los casos problemáticos, si los hubo, que no se atendieron en algún momento.

Las tareas aquí se dirigen, en definitiva, hacia una lectura detenida y exhaustiva de toda la información contenida en cada una de las entradas con el fin de detectar y subsanar errores (campos vacíos, sintaxis apropiada en contextos y definiciones, ortografía, etc.):

| *TAREAS* | | | *revisión* | *observaciones* |
|---|---|---|---|---|
| 1. | Repetición de registros | | | |
| 2. | Campos obligatorios completados | | | |
| 3. | Funcionamiento correcto de referencias cruzadas | | | |
| 4. | Pertinencia de los términos | | | |
| 5. | Revisión ortográfica y gramatical | | | |
| 6. | Sustantivos (en singular o plurales lexicalizados) | | | |
| | Verbos (en infinitivo) | | | |
| | Adjetivos (con/sin flexión de género) | | | |
| 7. | Redacción de definiciones | | | |
| 8. | Contextos | Sintaxis | | |
| | | Contenido (reproducen exactamente la unidad terminológica) | | |
| 9. | Referencias bibliográficas | | | |
| 10. | Equivalente: misma categoría gramatical o misma función gramatical | | | |

**Tabla 5:** Tareas de la fase de revisión

Una vez incorporadas todas las modificaciones y correcciones que se consideran pertinentes, sólo queda elaborar una versión definitiva del trabajo.

### *2.3.5 Edición*

La última fase de un proyecto terminológico es la edición (impresa o digital) de la aplicación que se pondrá a disposición del usuario final. Es claro que dependiendo del soporte elegido en un principio las tareas que se tendrán que acometer serán distintas. De todas formas, el producto elaborado requerirá cierto grado de adaptación. En el caso de un diccionario que va a ser publicado en papel por terceros, será necesario volcar los datos electrónicos contenidos en la base de datos a un formato de texto (.doc, .rtf, etc.) para que puedan manipularse en la etapa de maquetación. Si,

por el contrario, va a publicarse en formato electrónico (CD-ROM, Internet, etc.), bien podemos contactar con el servicio técnico del SGBDT elegido para averiguar las posibilidades, o bien utilizar los programas necesarios destinados a tal fin.

## 3. Conclusiones

El objetivo de este trabajo ha sido presentar la organización y sistematización del trabajo terminográfico bilingüe que se realiza dentro del marco del grupo de investigación *IPA*, útil para la elaboración de diccionarios especializados bilingües destinados al traductor. Para alcanzarlo, hemos consagrado el presente artículo a la exposición resumida del método que nos permite sistematizar y automatizar la gestión terminológica bilingüe desde el inicio del proyecto terminológico hasta la publicación del conjunto estructurado de términos del ámbito objeto de estudio. Se han descrito las etapas del proceso terminográfico que se desarrollan en el marco investigador aludido y se han definido cada una de las tareas que se llevan a cabo en dichas etapas, junto con las herramientas informáticas empleadas, a excepción de algunas tareas específicas que se realizan con los corpus, por estar ya descritas en trabajos anteriores (*cf.* Vargas, 2006a y b). Todas las fases mencionadas, junto con las tareas que se llevan a cabo en cada una de ellas y las herramientas que hacen posible su automatización constituyen nuestra propuesta de sistematización global de la terminografía descriptiva para la elaboración de diccionarios traductológicos bilingües (inglés-español).

## Bibliografía


AARTS, B. (1991): «Intuition-based and observation-based grammars». En Aijimer, K. y Altenberg, B. (eds.): *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London: Longman, pp.44-62.

AHMAD, K., Davies, A., Fulford, H. y Rogers, M. (1994): «What is a term? The semi-automatic extraction of terms from text». En SNELL-HORNBY, M., PÖCHHACKER, F. y KAINDL, K. (eds.): *Translation Studies: An Interdiscipline*. Amsterdam/Philadelphia: Jon Benjamins, pp.267-278.

ARNTZ, R. y Picht, H. (1989): *Einführung in die Terminilogiearbeit*. Hildesheim: Georg Olms Verlag. [Trad. *Introducción a la terminología*. Madrid: Pirámide, 1995]

ATKINS, B.T.S. Clear, J. Y Ostler, N. (1992): «Corpus Design Criteria». En *Literary and Linguistic Computing*, vol.7, n. 1, 1996, pp.1-16.

BOWKER, L. (1996): «Towards a Corpus-Based Approach to Terminography». En *Terminology*, 3(1), 1996, pp. 27-52.

CABRÉ, M.T. (2004): "De los diccionarios a los bancos de conocimiento: nuevas herramientas del traductor". En *El español, lengua de traducción. II congreso internacional*. Bruselas*:* ESLEtRA*,* págs. 23-55. Disponible en: http://www.upf.edu/pdi/dtf/teresa.cabre/docums/ca04ban.pdf. [Última fecha de consulta: agosto 2008].

CLEAR, J. (1993): «From Firth Principles. Computational Tools for the Study of Collocation». En Baker, M., Francis, G. y Tognini-Bonelli, E. (eds.): *Text and Technology – In Honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins, pp. 271-292.

EAGLES. 1995. *Evaluation of Natural Language Processing Systems,* EAGLES document EAG-EWG-PR.2. Version of September, 1995. Disponible en línea: http://www.issco.unige.ch/ewg95/ewg95.html [Última fecha de consulta: agosto 2008].

GTW - Gesellschaft fürTerminologie und Wissenstransfer e.V. (1996). «Criteria for the Evaluation of Terminology Management Software». In: GTW Report. Saarbrücken: GTW.

HALLIDAY, M.A.K. (1991): «Corpus studies and probabilistic grammar». En Aijmer, K y Altenberg, B. (eds.): *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London/New York: Longman, pp.30-43.

ISO (12620, 1999E): *Computer applications in terminology —Data categories*.

ISO (16642, 2003E): *Computer applications in terminology —Terminological markup framework*.

IULA. «Metodología de Trabajo en Terminología» [en línea]. En *Grup IulaTerm. Tallers online de Terminologia*. Barcelona: IULA. Universidad Pompeu Fabra, 2006. <http://www.iula.upf.edu/iulonlca.htm>.

LEECH, G. (1991): «The state of the art in corpus linguistics». En Aijmer K. y Altenberg B. (eds.): *English Corpus Linguistics: Studies in Honour of Jan Svartvik, London: Longman*, pp 8-29. Disponible en línea: <http://angli02.kgw.tu-berlin.de/corpus/art.htm>. [Última fecha de consulta: agosto de 2008].

Lorente, M. (2001): «Teoría e innovación en terminografía: la definición terminográfica». En Cabré, M. T. y Feliu, J. (eds.): *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica (DGES PB96-0293)*, pp. 81-112.

PAVEL, S. y Nolet, D. (2001): *Manual de terminología*. Canadá: Ministre des Travaux publics et Services gouvernementaux. Disponible en línea: <http://www.bureaudelatraduction.gc.ca/publications/documents/termino-spa.pdf>. [Última fecha de consulta: agosto 2008]

PEARSON, J. (1998): *Terms in Context*. Amsterdam/Philadelphia: John Benjamins.

ROBERTS, R. P. y Langlois, L. (2001): «L'apport de l'informatique à la recherche lexicographique». En *Meta*, 46(4), pp. 711-720.

SECO, M., Andrés, O. y Ramos, G. (1999): *Diccionario del Español Actual*. Madrid: Aguilar Lexicografía.

Sinclair, J. (1991): *Corpus, Concordance, Collocation*. Oxford: OUP.

Sinclair, J. (2003): «Corpora for lexicography». En Sterkenburg, P. (ed.) (2003): *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: John Benjamins, pp.167-178.

TERMCAT (1990): *Metodologia del treball terminològic*. Barcelona: Generalitat de Catalunya, Departament de Cultura.

VARGAS SIERRA, C. (2005a): «A pragmatic model of text classification for the compilation of special-purpose corpora». En MATEO, J. y YUS, F. (eds.): *Thistles. A homage to Brian Hughes. Essays in Memorian* (vol. II), pp. 295-315. Disponible en línea: http://www.ua.es/personal/chelo.vargas/Documentos/BH%20Vargas.pdf. [Última fecha de consulta: agosto 2008].

VARGAS SIERRA, C. (2005b): *Aproximación terminográfica al lenguaje de la piedra natural. Propuesta de sistematización para la elaboración de un diccionario traductológico*. Tesis doctoral. Alicante: Universidad de Alicante.

VARGAS SIERRA, C. (2006a): «Diseño de un corpus especializado con fines terminográficos: el Corpus de la Piedra Natural». En *Debate Terminológico*, 2 (7/2006). París: RITERM (Red Iberoamericana de Terminología). Disponible en línea: http://www.riterm.net/revista/n_2/chelo_vargas_sierra.pdf [Última fecha de consulta: agosto 2008].

VARGAS SIERRA, C. (2006b): «El proceso terminográfico multilingüe con WordSmith Tools». En *CONFLUENCIAS - Revista de Tradução Científica e Técnica*, n.4, pp. 84-107. Disponible en línea: http://www.confluencias.net/n4/vargas-sierra.pdf. [Última fecha de consulta: agosto 2008].

Anexo : Control de documentos

| | |
|---|---|
| Identificador asignado por la BD: | |
| Nombre asignado al documento: | |
| Procesador: | |

Módulo I de Identificación

| | |
|---|---|
| Título: | |
| ☐ Formato electrónico | |
| Formato del documento: | |
| Nombre del documento original: | |
| Origen ☐ Internet | |
| Motor de búsqueda: | |
| Sintaxis: | |
| URL: | |
| Fecha de consulta: | |
| Comentarios: | |
| Origen ☐ CD-Rom ☐ Expertos ☐ Otros | |
| Comentarios: | |
| ☐ Formato papel | |
| Publicado en: | |
| Lugar de edición: | |
| Fecha OCR: | |
| Fecha fin de revisión: | |
| Estado del copyright: | ☐ Sin iniciar |
| | ☐ Iniciado |
| | ☐ Rechazado |
| | ☐ Resuelto |
| Nº. de palabras: | |
| Fecha de recuperación: | |
| Nº. de personas que crean el texto | ☐ varias |
| | ☐ una |
| | ☐ desconocido |
| Nombre del autor/es: | |
| Año de publicación: | |
| Comentarios: | |

Módulo II (lingüístico)

| | |
|---|---|
| Lengua del documento: | |
| Estatus lingüístico: | ☐ Original ☐ Traducción |
| Fiabilidad: | ☐ Alta ☐ Baja ☐ Media ☐ Ninguna |
| Lengua Materna: | |

| Documentos paralelos: | Lengua | Número |
|---|---|---|
| | | |
| | | |
| | | |
| | | |

| |
|---|
| Nombre asignado al documento: |

Módulo III (especialización y género)

- ☐ Experto – Experto
  - ☐ Didáctico – Instructivo
    - ☐ manuales de instrucciones
    - ☐ instrucciones de trabajo
    - ☐ plan de producción
  - ☐ Informativo
    - ☐ catálogo de fichas técnicas
    - ☐ informe técnico
    - ☐ artículo de revista especializada
    - ☐ artículo especializado comercial
    - ☐ artículo de revista sectorial
    - ☐ monografía
    - ☐ anuario
    - ☐ informe de proyecto
  - ☐ Jurídico normativos
    - ☐ patente
    - ☐ norma laboral
    - ☐ norma de ensayo
  - ☐ Recopilatorio
    - ☐ diccionarios
    - ☐ vocabulario
    - ☐ glosario

- ☐ Experto - Semiexperto
  - ☐ Didáctico - Instructivo
    - ☐ manual de instrucciones
    - ☐ manual técnico
    - ☐ libro de texto
  - ☐ Informativo
    - ☐ artículo de revista especializada
    - ☐ artículo de revista sectorial
    - ☐ anuario
    - ☐ catálogo de fichas
    - ☐ monografía
  - ☐ Recopilatorio
    - ☐ glosario
    - ☐ artículo enciclopédico
    - ☐ vocabulario
    - ☐ diccionario

- ☐ Experto - Lego
  - ☐ Divulgativo
    - ☐ folleto publicitario
    - ☐ artículo divulgativo
    - ☐ catálogo
  - ☐ Recopilatorio
    - ☐ artículo enciclopédico
    - ☐ diccionario
    - ☐ glosario
    - ☐ vocabulario