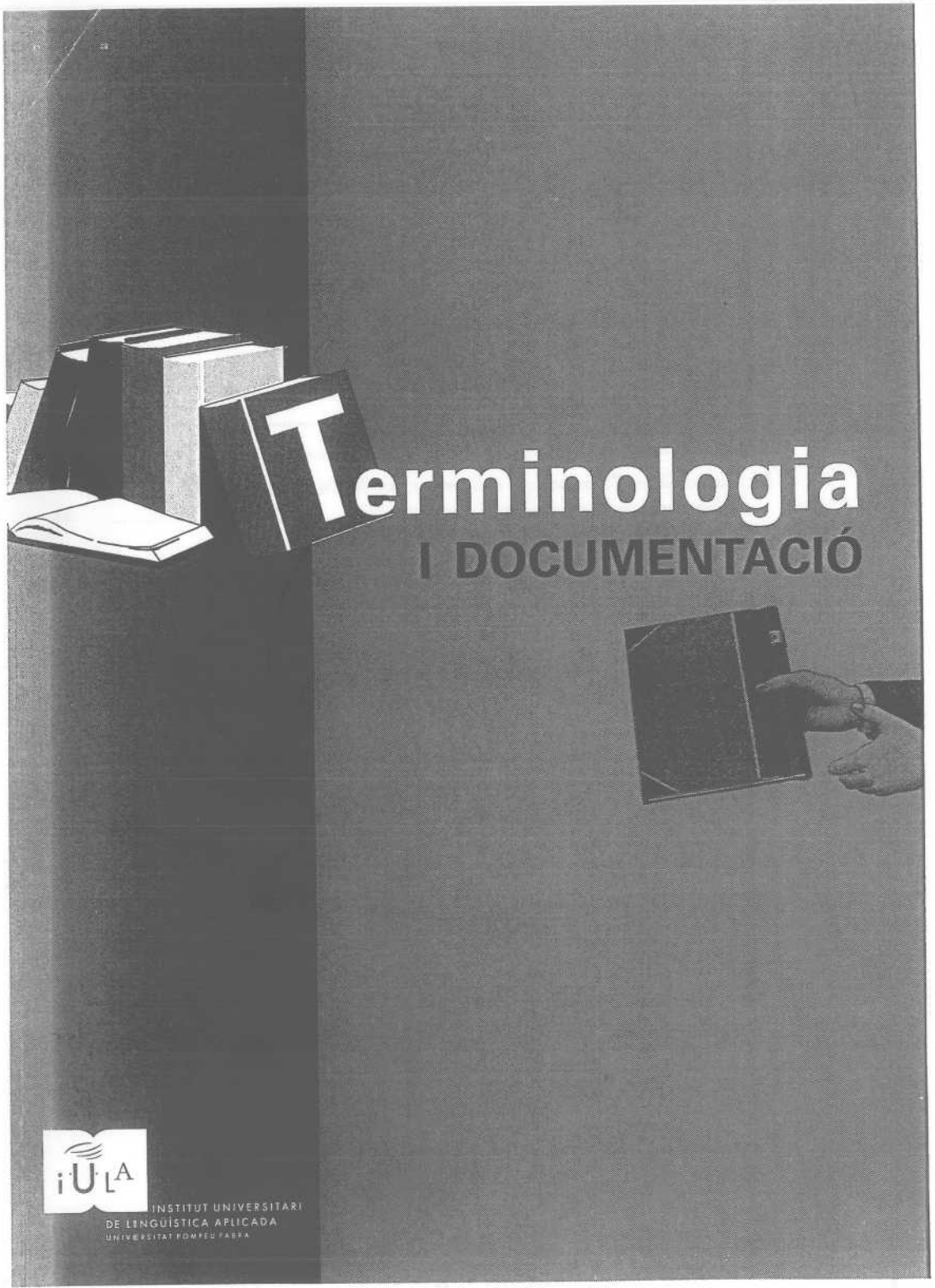
Terminologia
I DOCUMENTACIÓ
iULA
INSTITUT UNIVERSITARI
DE LINGÜÍSTICA APLICADA
UNIVERSITAT POMPEU FABRA

BIBLIOTECA DE CATALUNYA. DADES CIP

Jornada de Terminologia i Documentació *(1a : 2000)*
Terminologia i documentació : I Jornada de Terminologia i Documentació (24 de maig de 2000). - (Sèrie activitats)
Textos en català i castellà. - Bibliografia
ISBN 84-477-0734-2
I. Cabré, M. Teresa, ed. II. Codina, Lluís, ed. III. Estopà, Rosa, ed. IV. Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada V. Títol VI. Col·lecció: Sèrie activitats (Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada)
1. Terminologia - Congressos 2. Recuperació de la informació - Congressos 3. Tesaurus - Congressos
800.3:025.4 (061.3)

Responsables de l'edició: M. Teresa Cabré, Lluís Codina i Rosa Estopà
Tasques de maquetació: Yannick Garcia

Direcció de les Publicacions de l'IULA: M. Teresa Cabré
Coordinació de les Publicacions de l'IULA: Mercè Lorente, Gemma Martínez

Primera edició: març de 2001

La Rambla, 30-32
08002 Barcelona

Disseny de la coberta: Cass
Impressió: Catalana de Formularis, S.L.
Dipòsit legal: B-957-2001
ISBN: 84-477-0734-2

ÍNDEX

Automatización de tareas en la elaboración de un diccionario terminológico*

Mª Amparo Alcina Caudet
Departament de Traducció i Comunicació
Universitat Jaume I

1. Introducción

La presente comunicación pretende ser una reflexión sobre la metodología del trabajo terminológico y el modo en que podemos utilizar los ordenadores en algunas fases del proceso. Estas reflexiones surgen a partir del desarrollo de un proyecto de terminología que actualmente llevamos a cabo un equipo de investigadores del Departamento de Traducción y Comunicación de la Universitat Jaume I, cuyo objetivo es el estudio del lenguaje de especialidad y la terminología de la cerámica.

Durante la elaboración del diccionario terminológico, nos ha servido de gran ayuda el uso de un programa gestor de bases de datos terminológicas, concretamente Multiterm de Trados, que nos ha facilitado el trabajo fundamentalmente en la fase de recopilación de la terminología y presentación de los datos obtenidos. Sin embargo, en otras fases del trabajo terminológico no hemos podido disponer de herramientas que nos facilitaran el manejo de los datos o bien hemos utilizado programas informáticos que no permitían la interacción, por lo que los resultados obtenidos tenían que ser transferidos de un modo «semi-automático» o manual.

En mi opinión, el puesto de trabajo del terminólogo puede mejorar cualitativamente creando sistemas que añadan a las actuales ventajas de los bancos de datos terminológicos otra serie de herramientas o funciones para llevar a cabo las distintas tareas que forman parte del trabajo terminológico. La investigación en informática y los cada vez más sofisticados productos de hardware han permitido aumentar la rapidez, la capacidad de almacenamiento de datos, la conversión de información a soportes electrónicos. Por otra parte, se ha incrementado notablemente la cantidad y calidad de programas orientados al tratamiento lingüístico de textos. Algu-

* Esta comunicación tiene como marco el proyecto de investigación *Estudio y descripción de la lengua de especialidad de una rama profesional de la cerámica y elaboración de un diccionario terminológico multilingüe*, desarrollado en el Departamento de Traducción y Comunicación de la Universitat Jaume I (*http://www1.uji.es/wwwtrad*) y financiado por la Fundació Caixa de Castelló - Bancaixa (referencia P1A98-12).

nos de estos programas pueden ser perfeccionados o adaptados al trabajo terminológico.

Por ello, pretendo aquí en primer lugar analizar la aplicación que pueden tener en terminología algunas herramientas de uso lingüístico o genérico; en segundo lugar, plantearé un modelo de programa informático que tenga en cuenta la interacción entre estas herramientas.

2. Terminología asistida por ordenador

En los últimos 20 años se han creado programas comerciales o gratuitos que, aunque no fueron diseñados estrictamente para la elaboración de terminologías, permiten automatizar algunas de las tareas relacionadas con la creación de diccionarios especializados. Algunos de estos programas son genéricos, de modo que se pueden utilizar para gran número de fines. Por ejemplo, las bases de datos se pueden utilizar en el contexto de la terminología y la documentación para guardar los datos de los libros de una biblioteca; los programas de hipertexto se pueden utilizar para presentar informaciones de muy diversos tipos; las bases de conocimiento se utilizan para el diseño de programas informáticos y sistemas expertos.

Otros programas son específicamente para uso lingüístico, por ejemplo, los programas de concordancias, los programas de análisis textual, las bases de datos terminológicas.

2.1. *Programas para elaboración de terminologías*

1) Bases de datos genéricas y bases de datos bibliográficas

En las bases de datos podremos recopilar la información de las fuentes bibliográficas interesantes bien para documentarnos sobre el tema del cual queremos hacer el diccionario, bien para utilizar como corpus de vaciado, bien para localizar las equivalencias en otras lenguas.

2) Programas de concordancias

Son programas que toman textos en formato ASCII (lo que conocemos en Windows como formato de texto TXT) y son capaces de realizar búsquedas, comparaciones y cálculos entre las cadenas de caracteres que aparecen en él. Están preparados para realizar cierto tipo de tareas, como:

a) listados de las palabras que aparecen en el texto ordenadas de modos diversos (por ejemplo: por orden de frecuencia, por orden alfabético normal o inverso, según la longitud de la palabra, etc.);

b) índice de palabras, o listado de palabras, en el que se indica junto a cada palabra el número de veces que aparece en el texto, y referencias indicando su ubicación en el texto;

c) concordancias de una determinada palabra, es decir, listados de los contextos en los que aparece una determinada palabra en el texto;

d) tablas de colocaciones: en las que se nos muestra con qué frecuencia una determinada palabra aparece junto a otras;

e) estadísticas: el programa puede también proporcionarnos datos estadísticos sobre el número absoluto de palabras del texto, el número de palabras diferentes que aparecen en el texto, cálculos de frecuencia, etc.

Dado que se trata de programas que están preparados para trabajar con cadenas de caracteres alfanuméricos, los usos y aplicaciones están especialmente dirigidos hacia los estudios filológicos y lingüísticos. Han sido utilizados por los filólogos para estudiar las obras literarias (uso del vocabulario por ciertos autores), y también por los lingüistas para realizar estudios lexicográficos, diccionarios inversos, diccionarios de frecuencias. Como ejemplos de este tipo de herramientas podemos citar: MonoConc y WordSmith.

En la elaboración de diccionarios terminológicos, estos programas nos pueden ayudar a hacer una primera tentativa de análisis lexicográfico de los textos del corpus de vaciado. Podemos pedir un listado de las palabras del texto y pedir que se eliminen de ese listado las palabras gramaticales. Además, una vez concluido el vaciado terminológico, podremos comparar la terminología extraída con el listado de palabras obtenido de forma automática para comprobar posibles ausencias de términos.

3) Programas de análisis textual

Se trata de utilidades que permiten manejar textos anotados. Las palabras que aparecen en estos textos contienen etiquetas que describen sus características morfológicas, léxicas, sintácticas o discursivas. El investigador puede introducir en el texto las anotaciones que considere convenientes en función de los parámetros que pretenda investigar. Es necesario determinar, en primer lugar, qué tipo de información es necesario codificar en el texto y, en segundo lugar, qué etiquetas pueden representar este tipo de información. Existen diferentes formas de configurar la estructura interna de las anotaciones, que dan lugar a distintos sistemas de

marcaje, entre los que se encuentran, por ejemplo, COCOA y SGML. En COCOA, por ejemplo, el formato de anotación consiste en una estructura como la siguiente: <etiqueta valor>, donde *etiqueta* será un código (por ejemplo T, A) que representará algún tipo de información (como tipo de texto, autor) y *valor* será la información que para esa etiqueta recibe ese texto o fragmento del corpus. Por ejemplo, la etiqueta <A Carmen Alborch> indicará que Carmen Alborch es la autora del texto. Una vez introducida una anotación COCOA, la información que aporta al texto afecta a todo el texto que aparezca a continuación mientras no se introduzca una nueva anotación con la misma etiqueta y distinto valor.

El sistema de marcaje SGML es más complejo. Utiliza cuatro conceptos básicos: la entidad de marcaje, el elemento de marcaje, los atributos del elemento de marcaje y el tipo de documento. La entidad de marcaje es un objeto concreto del texto: un conjunto de caracteres, un gráfico, un título, etc. El elemento de marcaje es la etiqueta o representación que utilizamos para señalar un tipo de información, por ejemplo, *name* serviría para señalar la información «nombre propio». Los elementos de marcaje están delimitados por los caracteres '<' y '>', que codifican la etiqueta de apertura, y '</' y '>', que codifican la etiqueta de cierre. Entre la etiqueta de apertura y la de cierre deberá aparecer la entidad de marcaje. Los atributos de un elemento de marcaje introducen información adicional relacionada con un elemento de marcaje; constan de un nombre y de un valor. Por ejemplo, las etiquetas <pb> y </pb> sirven para delimitar el principio y el fin de una página. A la etiqueta de apertura podemos añadir el atributo *n* para indicar el número de página y tendremos: <pb n=6> texto de la página </pb>. Por último, el tipo de documento es la especificación de todos los elementos de marcaje, los atributos que tienen asociados y los valores posibles para cada uno de ellos, en definitiva, la gramática utilizada para codificar el texto. La especificación de estos códigos y valores, permitirá que el programa pueda interpretar correctamente el texto y también le permitirá detectar posibles errores que se hayan podido producir en la etiquetación del texto.[1]

Como ejemplos de programas que pueden manejar corpus etiquetados están LEXA, Micro-OCP, TUSTEP, WordCruncher y TACT.

La incorporación de este tipo de herramientas al trabajo terminológico nos permitirá extraer de los corpora no sólo las denominaciones terminológicas, con sus correspondientes categorías gramaticales, colocaciones, frecuencias, etc. También podremos anotar en estos textos la información que nos interesaría extraer en la

[1] Pérez (1998) ofrece una excelente exposición de los sistemas de codificación, la creación de corpus anotados y los programas que facilitan el trabajo con corpus electrónicos.

fase de vaciado: contextos definitorios, contextos lingüísticos, contextos de equivalencias. Una vez anotado el corpus con esta información, podríamos pedir al programa una relación de los contextos definitorios de un determinado término para poder, a partir de ellos, elaborar su definición terminológica.

4) Gestores de bases de datos terminológicas

Estos programas nos permiten recopilar distintas informaciones relacionadas con los términos. La información se estructura en campos y registros o fichas. El usuario puede definir la estructura de campos según el tipo de información que le interese recopilar (por ejemplo, categoría gramatical, definición, ejemplos de uso, área temática, etc.). Una vez diseñada la estructura de campos, se introducirán en cada ficha los datos que se refieren a una misma unidad de información, por ejemplo, un término o un concepto.

Los sistemas de gestión de bases de datos terminológicas permiten funciones como la consulta automática de los términos, la búsqueda de un conjunto de fichas que contienen una determinada información en alguno de sus campos, la ordenación de las fichas en función de algún criterio, la modificación de una ficha (añadir, eliminar o cambiar los datos), la exportación de datos a otros tipos de formato o la importación de datos de otros formatos, y la difusión de la terminología (bien en soporte papel o electrónico).

Ejemplos de bases de datos terminológicas son Termex, Multiterm, DicTip.

Aunque los sistemas de gestión terminológica están principalmente orientados a presentar ficheros terminológicos y facilitar su difusión, se pueden usar también en la fase de elaboración de terminologías sistemáticas. Se trata de definir en ellos campos como contexto definitorio, contexto lingüístico, contexto de equivalencia y llevar a cabo el vaciado de los textos del corpus en estos sistemas para elaborar el fichero terminológico, en lugar de usar las tradicionales fichas de vaciado en papel. Una vez terminada la fase de vaciado, en cada ficha tendremos recogidos todos los contextos en que ha aparecido un determinado término, y que nos servirán para elaborar las definiciones, para proporcionar ejemplos de uso y para decidir las relaciones de sinonimia, área temática, etc.

5) Gestores de bases de conocimiento

Un gestor de bases de conocimiento es un programa que recopila y gestiona información conceptual. Permiten estructurar las características que pertenecen a un concepto y establecer distintos tipos de relaciones entre conceptos, del tipo «género - especie», «tipo de», «parte de». Además, algunas relaciones se pueden determi-

nar automáticamente, por ejemplo, la herencia. Si un determinado concepto específico, *mosaico*, es un tipo de *baldosa*, recibirá automáticamente las características que contenga el concepto *baldosa*.

Por otra parte, permiten representar gráficamente estas relaciones (por ejemplo, en forma de tablas, diagramas o árboles de dependencias) de modo que resulta más evidente e intuitiva la interpretación y comprensión del sistema de conceptos.

Puesto que trabaja con los términos en función de sus características conceptuales y los organiza no como simples listados de palabras, sino como sistemas de conceptos, ofrece muchas ventajas de cara a la metodología del trabajo terminológico: permite comparar los términos en función de las características que tienen en común y determinar en qué casos existen relaciones de sinonimia, comparar terminologías en distintas lenguas basándose en la similitud conceptual y, por último, facilita la labor de elaboración de definiciones.

El Laboratorio de análisis del lenguaje para la ingeniería del conocimiento de la Universidad de Ottawa[2] desarrolló el sistema de gestión del conocimiento CODE4 que ha sido utilizado para diseñar distintas terminologías (Meyer *et al.* 1997). Actualmente desarrollan el sistema DOCKMAN, que también incluye herramientas de análisis textual.

6) La integración de herramientas

He presentado una serie de programas que, bien por sí mismos, bien adaptados a las funciones requeridas por la labor terminológica, pueden desempeñar un gran papel en alguna de las fases del proceso de elaboración de terminologías. Sin embargo, la ausencia de interfaces entre estos programas obliga a crear complicados subprogramas o macros para que los resultados obtenidos en una fase del proceso puedan ser retomados en la fase siguiente.[3] En otros casos, la creación de esas macros resulta imposible y es necesario volver a teclear los resultados. En definitiva, la labor se hace compleja y a menudo se prefiere realizar manualmente ciertas tareas o improvisar procedimientos más rudimentarios mediante herramientas de uso genérico.

Resultaría pues, muy conveniente, disponer de un macroprograma o paquete integrado que reúna las funciones necesarias para la elaboración de terminologías y que proporcione las interfaces necesarias entre las distintas herramientas para poder aprovechar lo que cada una de ellas nos ofrece.

[2] Para más información, puede consultarse su página web *http://www.site.uottawa.ca/lake/index.html.*

[3] Calzolari (1997) examina las posibilidades de interacción entre corpus y bases de datos léxicas.

El diseño de este paquete integrado constaría de las herramientas siguientes:

a) gestor de fichero de fuente;

b) programa de análisis textual;

c) gestor de base de datos terminológica;

d) gestor de base de conocimiento.

El fichero de fuentes contendrá las referencias bibliográficas de las obras que van a ser utilizadas como corpus de vaciado o para buscar las equivalencias. El fichero de fuentes debe poder ser consultado desde el corpus etiquetado y desde la base de datos terminológica para obtener la referencia completa de las obras que se mencionan.

El programa de análisis textual permitirá tomar el corpus electrónico y anotarlo con etiquetas que permitirán posteriormente la extracción de información relevante desde el punto de vista terminológico: denominación, categoría gramatical, contexto definitorio, contexto lingüístico, contexto de equivalencias, notas, fuente, área temática. Deberá, asimismo, contener una interfaz con el gestor de base de datos terminológica en la que se trasvasen del corpus a la ficha terminológica estas informaciones estructuradas.

Una vez extraída la información del corpus, el terminólogo dispondrá de una base de datos terminológica a partir de la cual podrá analizar los distintos términos, extraer características y establecer las relaciones entre conceptos con ayuda del gestor de la base de conocimiento.

El gestor de la base de conocimiento facilitará la elaboración semi-automática de definiciones. Estas definiciones serán posteriormente copiadas a la base de datos terminológica. Por otra parte, la información conceptual podrá servir para etiquetar semánticamente el corpus.

En la Figura 1 se muestra un esquema de las herramientas que debería contener este programa y de las interrelaciones entre ellas.

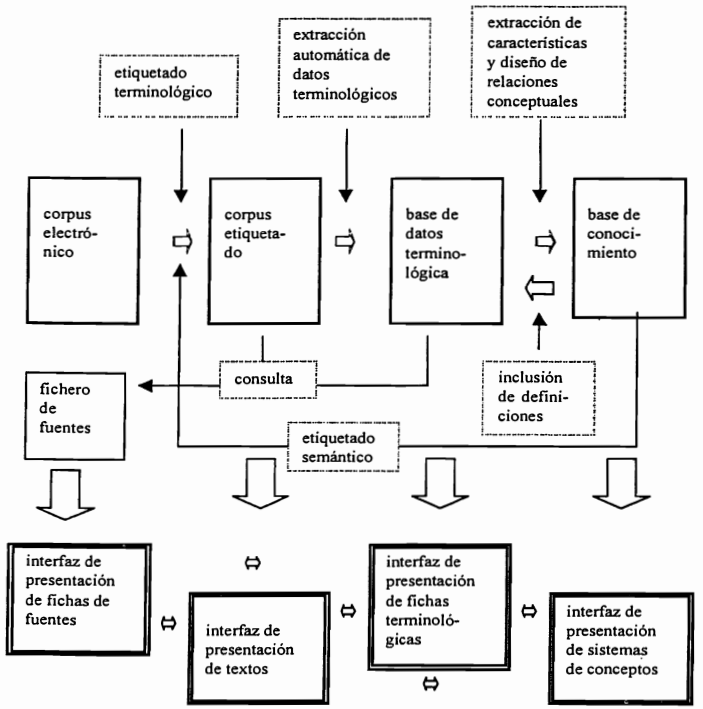

Figura 1. Esquema de un paquete integrado de terminología asistida por ordenador

4. Conclusiones

La sociedad de la información demanda recursos terminológicos para un acceso rápido y eficaz a la información y a la comunicación plurilingüe. Para hacer frente a este reto, debemos rentabilizar nuestros esfuerzos y automatizar las tareas que entran en juego en el trabajo terminológico.

Un paquete integrado que reúna estas características no sólo facilitará la elaboración de terminologías, sino que también facilitará la difusión a los distintos sectores interesados (especialistas en la materia de que se trate, traductores, documentalistas), mejorará la calidad de las terminologías (al incorporar no sólo listados de palabras y definiciones, sino también las relaciones conceptuales que existen entre los términos), y permitirá su reutilización en otros sistemas de inteligencia artificial que requieran este tipo de productos (como la lingüística del corpus, la traducción automática y los sistemas expertos).

Bibliografía

Calzolari, N. (1997). «Lexicon and Corpus: a multi-faceted interaction». Dins Cabré, M. T. (dir.) (1997). *Cicle de conferències 95-96. Lèxic, corpus i diccionaris*. Barcelona: Institut Universitari de Lingüística Aplicada - Universitat Pompeu Fabra. 77-89.

Meyer, I.; Eck, K.; Skuce, D. (1997). «Systematic concept analysis within a knowledge-based approach to terminology». Dins Wright, S. E.; Budin, G. (ed.) (1997). *Handbook of Terminology Management*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

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