EXTRACCIÓN DE TERMINOLOGÍA: ELEMENTOS PARA LA CONSTRUCCIÓN DE UN EXTRACTOR

Rosa Estopà Bagot[*]

RESUMEN: Esta presentación pretende exponer algunas de las principales conclusiones de nuestra tesis doctoral Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Terminologia), presentada en julio de 1999 en la Universidad Pompeu Fabra y dirigida por la Dra. M. Teresa Cabré. El texto se articula en cuatro puntos. En el primero, se presentan los resultados del análisis crítico de los principales extractores de terminología existentes. En el segundo y tercer apartados, se discuten las causas de las limitaciones de estos sistemas. Finalmente, en el cuarto punto, se presentan dos parámetros para mejorar un sistema de extracción de terminología: la ampliación del objeto de extracción y la funcionalidad de la extracción.

PALABRAS-CLAVE: terminología, extracción de terminología, aplicaciones terminológicas.

RESUMO: *Esta apresentação pretende expor algumas das principais conclusões da nossa tese intitulada Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Terminologia), apresentada em julho de 1999 na Universidade Pompeu Fabra e orientada pela Dra. M. Teresa Cabré. O artigo articula-se em quatro pontos. No primeiro, serão apresentados os resultados da análise dos principais extratores de terminologia existentes. No segundo e terceiro apartados,*

[*] Facultad de Traducción y Interpretación e Instituto Universitario de Lingüística Aplicada de la Universidad Pompeu Fabra, Barcelona, España.

serão discutidas as causas das limitações destes extratores. Finalmente, no quarto ponto, serão apresentados dois dos parâmetros que propusemos na tese para melhorar um extrator: a ampliação do objeto de extração e o ponto de vista funcional da extração.

UNITERMOS: *terminologia; extração de terminologia; aplicações terminologicas.*

El reconocimiento de las unidades de un texto con significado especializado conocido como vaciado terminológico es una de las fases básicas de todo trabajo en terminología (elaboración de vocabularios, glosarios, bases de datos, bases de conocimiento, tesauros, preparación de traducciones, indización de textos, construcción de correctores ortográficos, etc.). Pero si bien es una tarea central, al mismo tiempo no es una tarea nada simple, sino que requiere mucho tiempo y sistematicidad en la aplicación de criterios. Es, de hecho, una de las fases más fatigosas y largas –sobre todo cuando se manipulan volúmenes de información importantes–que tiene el riesgo de convertirse en poco sistemática y, por consiguiente, ineficaz. Con la función de automatizar la fase de vaciado del trabajo terminológico para ganar rapidez y sistematicidad, a finales de la década de los ochenta, se concibió el primer extractor automático de terminología (TERMINO, 1988).

Un extractor de terminología podría definirse como un conjunto de programas informáticos que reconoce y extrae las unidades terminológicas (UT) que aparecen en un corpus de textos especializados.

En esta sesión nos proponemos plantear algunas cuestiones en relación con los extractores de terminología que fueron estudiadas en nuestra tesis doctoral; tesis que tenía como finalidad aplicada proponer elementos para mejorar el diseño de un extractor.

1. Los sistemas de extracción automática de terminología

Antes pero de proponer las bases de un nuevo extractor, tuvimos que analizar y evaluar el funcionamiento de los principales

sistemas de extracción automática de candidatos a términos existentes. En total, estudiamos 18 extractores de distinta naturaleza, basados en estrategias metodológicas lingüísticas, estadísticas o mixtas[1] (Estopà, Vivaldi, Cabré, 1998).

Para este trabajo preliminar establecimos seis parámetros de análisis que caracterizasen los extractores: la información de partida, las estrategias de reconocimiento, las estrategias de filtrado, las estrategias de adquisición del conocimiento, la interacción del sistema con el usuario y finalmente los resultados que obtenían, es decir el grado de satisfacción.

Del análisis de estos parámetros deducimos las siguientes conclusiones:

1. Todos los sistemas de extracción de unidades terminológicas analizados:
   - 1.1 generan demasiado silencio, es decir las unidades que un extractor debería reconocer y no reconoce, y demasiado ruido, es decir las unidades que el extractor ha seleccionado como candidatos y no debería haber reconocido.
   - 1.2 proponen demasiados candidatos a término que el usuario debe validar manualmente
   - 1.3 se centran exclusivamente en la unidad terminológica entendida como una unidad nominal

[1] Los extractores estudiados fueron los siguientes: ACABIT [Daille, 1994]; ANA [Enguehard et Pantera, 1994]; ATELIER/FX [www.ling.uqam.ca/Ato/Fx/AtelierFX.html]; AUTOLEX [Planas, 1994]; BLANK [Blank, 1995]; CLARIT [Evans et Zhai, 1996]; DROUIN [Drouin, 1997]; FASTR [Jacquemin, 1996]; HEID [Heid et alo., 1996]; LEXTER [Bourigault, 1994]; NAULLEAU [Naulleua, 1998]; NEURAL [Frantzi et Ananiadou, 1995]; NODALIDA-95 [Arppe, 1995]; SBIC [Anzaldi, 1996]; TERMIGHT [Dagan et Church, 1994]; TERMINO [David et Plante, 1991]; TERMS [Justeson et Katz, 1995]; STELLA [Jacquin et Liscouet, 1996].

2. La mayoría de los sistemas
   - 2.1 se aplican a una sola lengua, que suele ser el inglés o el francés. A principios de 1999 no existe un extractor para el catalán ni para el castellano.
   - 2.2 se basan en la extracción de un tipo de unidad terminológica: la unidad terminológica poliléxica
   - 2.3 utilizan patrones para detectar las unidades
   - 2.4 se centran en una misma y única técnica para detectar y extraer todas las unidades
3. Ninguno de los extractores:
   - 3.1 no utilizan información semántica, a excepción del extractor de NAULLEAU (1998)
   - 3.2 no discriminan entre unidades terminológicas y otros tipos de unidades como las fraseológicas
   - 3.3 no utilizan en profundidad las características combinatorias y contextuales de los términos.

En consecuencia, delante de un texto como el siguiente, la mayoría de los extractores estudiados sólo detectarían las unidades remarcadas:

> Diagnóstico R. conorii requiere, para su cultivo, la inoculación en animales (cobai) o la **utilización de células VERO** o fibroblastos L.929. Mediante **centrifugación-Shell Vial** se puede detectar R. conorii en **muestras de sangre** en menos de 72 h.
>
> En relación al **diagnóstico serológico**, parece demostrado que la mejor técnica para la disponibilidad, sensibilidad, especidifciddd y rapidez es la IFI. Esta técnica permite también investigar de **manera independiente** la **presencia de anticuerpos específicos IgM** con la **finalidad de distinguir** entre **infección actual** y **seropositividad residual.**

Pero si analizamos textos especializados, por ejemplo sobre enfermedades infecciosas, realizamos un vaciado manual de las unidades con significado especializado que contienen y después

comparamos estos vaciados con vaciados realizados automáticamente de los mismos textos, observamos que:

a. hay muchas unidades en el texto no seleccionadas por los extractores i, que, en cambio, transmiten un significado especializado: *diagnóstico, R. conorii, cultivo, inoculación, fibroblastos L. 929, sensibilidad, IFI,* etc.
b. hay unidades que los extractores seleccionan que no son unidades terminológicas: *utilización de células VERO, manera independiente, presencia de anticuerpos específicos IgM, finalidad de distinguir, infección actual,* etc.

Constatadas estas observaciones, parece lógico preguntarse ¿Por qué ocurren desajustes entre los vaciados manuales y los vaciados automáticos?, o planteado de otra manera ¿Por qué no funcionan satisfactoriamente los extractores de terminología?. La causa de esta insatisfacción reside en el objeto de extracción y las estrategias para detectar este objeto.

Una de las limitaciones de los extractores es que son muy restrictivos en relación al objeto, de manera que se suelen centrar en la detección de las UT poliléxicas (UTP), de categoría gramatical nominal, pues son las unidades más prototípicas y las más frecuentes de los textos especializados, y además son las que presentan características morfosintácticas más explícitas que facilitan su extracción.

Esta restricción ha implicado reducir el reconocimiento automático a un tipo de palabras: las UT y, dentro de esta categoría, a un único tipo de estructura: las unidades sintagmáticas. Los autores de extractores suelen justificar explícitamente esta restricción con argumentos como los siguientes:

1. La mayoría de las UT de un ámbito de especialización son unidades sintagmáticas y, por eso, es irrelevante complicar el sistema haciendo que reconozca los términos simples.
2. Formalmente, las unidades monoléxicas especializadas y las generales no se diferencian y, por eso, es imposible dis-

criminar entre las unidades monoléxicas especializadas y las generales.

3. Las unidades monoléxicas son mucho más polisémicas que las poliléxicas y, por eso, trabajar con heurísticas semánticas referidas a los términos monoléxicos es mucho más complicado que no introducir conocimiento semántico en las UTP.

En conjunto, es cierto que, desde un punto de vista morfosintáctico, las UTP son más fáciles de detectar que las unidades monoléxicas, dado que presentan una estructura morfosintáctica explícita que es controlable. Pero a pesar de ello, los tres argumentos se deben matizar.

En relación con el primer argumento, diversos autores afirman que alrededor del 80% de las terminologías están formadas por UTP. Ello se explica porque suelen basarse en las unidades codificadas en repertorios léxicos y terminológicos, pero si partiesen del uso real de los términos en los textos, estas afirmaciones ya no serían válidas. Pruebas que hemos realizado en el campo de la biomedicina, indican que el conjunto de unidades monoléxicas especializadas de un texto temáticamente especializado no se puede despreciar pues corresponde entre el 35% y el 45% de las unidades que transmiten un significado especializado.

Los tres argumentos también son discutibles por la relevancia categorial que se les atribuye. Así, si bien las UT, que son siempre nominales, son las USE más frecuentes de los textos especializados, los verbos, los adjetivos y los adverbios con un uso temáticamente especializado también son relevantes en estos textos.

Por eso, creemos que estas afirmaciones se podrían reconsiderar porque, aunque sea cierto que las unidades léxicas simples son bastante idiosincrásicas y muy polisémicas (y, consiguientemente, es difícil discriminar lingüísticamente cuando una unidad simple se utiliza con un sentido especializado o general), dentro de las unidades monoléxicas hay diferentes clases de palabras – derivadas, compuestas, abreviadas – que pre-

sentan algunas peculiaridades formales en las que los extractores se podrían basar para detectar gran parte de los términos monoléxicos.

Es consecuencia, pues, es parece necesario replantearse el objeto de base del trabajo en terminología y en concreto de los extractores, y revisar los intereses terminológicos reales de sus usuarios.

2. Limitaciones de los extractores: el silencio

Como decíamos si comparamos el vaciado manual hecho por un especialista y el vaciado automático a partir de patrones morfosintácticos de un mismo texto, observamos que éstos no coinciden.

Limitándonos inicialmente al objeto de extracción de los extractores clásicos (las unidades terminológicas poliléxicas), el análisis de las unidades marcadas manualmente permite diferenciar dos tipos de silencio relativos al objeto de vaciado del sistema de extracción automática: silencio intrínseco y silencio extrínseco.

Entendemos por silencio intrínseco el conjunto de segmentos especializados que no detecta un extractor y que debería detectar porque son unidades terminológicas poliléxicas (UTP). Y entendemos por silencio extrínseco el conjunto de unidades especializadas, que no son UTP, que un extractor ignora explícitamente porque no forman parte de sus objetivos de extracción. El especialista, en cambio, las señala como unidades especializadas pertinentes.

De las unidades marcadas por el especialista que el extractor utilizado (EXCAT1) no ha detectado, unas – las más abundantes–son unidades monoléxicas, y otras – las menos numerosas – son unidades poliléxicas. Desde la perspectiva de un extractor, el silencio intrínseco es involuntario, y el extrínseco, totalmente voluntario. Consiguientemente, desde el punto de vista de las finalidades de los sistemas tradicionales, sólo debemos valorar negativamente el silencio intrínseco. En cambio, desde la óptica del especialista, es mucho más significativo el silencio extrínseco que el intrínseco.

2.1 Silencio intrínseco

El silencio intrínseco afecta aproximadamente entre el 10% i el 5% de las UTP del texto. Las causas de este tipo de silencio son básicamente tres: errores de desambiguación. superposición de términos y términos escondidos discursivamente.

Los errores que se producen en la desambiguación morfosintáctica están condicionados por las características del desambiguador lingüístico y/o estadístico que se utilice. Cuanto más robusto sea el desambiguador menos errores habrá en el vaciado terminológico. Actualmente, los desambiguadores lingüísticos resuelven alrededor del 75% de las ocurrencias de un texto; el 25% restante se suele desambiguar mediante cálculos estadísticos. La mayoría de los desambiguadores estadísticos actuales generan sólo entre un 3% y un 5% de error del total de ocurrencias de un texto. Existen desambiguadores estadísticos muy potentes que llegan a generar sólo un 2% de error; en estos casos, pero, los corpus de entrenamiento suelen ser muy restrictivos temáticamente.

El porcentaje de error en la desambiguación de un texto puede comportar que una UTP no se reconozca porque esté etiquetada equivocadamente. Por ejemplo, en el corpus analizado, las UT *buit poplití*[2] *y immnunocomplexos circulants* son unidades silenciadas por EXCAT1 porque, en estos casos, los segmentos buit y immunocomplexos han sido procesados como adjetivos y en este texto son substantivos. De esta manera, las unidades *buit popliti y immnunocomplexos circulants*, que son UT formadas por un nombre y un adjetivo, se silencian ya que en el corpus etiquetado responden a la estructura A A que no es una estructura terminológica prevista.

Las unidades superpuestas, en cambio, son términos complejos, que a la vez contienen unidades simples o complejas que son también términos: *pneumonitis intersticial* es un UT, pero también lo *es pneumonitis; febre tacada d'Israel* es una UT, pero también lo son *febre tacada y febre*.

---

[2] Los ejemplos son de la lengua catalana pues para este trabajo se utilizaron mayoritariamente textos especializados en catalán.

En la detección automática de los términos superpuestos el problema surge cuando se quieren recuperar todas estas UT que forman parte de otras UT más complejas, dado que algunos constituyentes de las UTP pueden ser también UT aisladas.

Finalmente, el problema más complejo relacionado con el silencio intrínseco es, sin duda, detectar y extraer las unidades con significado especializado de las que, por razones discursivas, se ha suprimido el núcleo o el complemento (o una parte del complemento) y los componentes que quedan se coordinan con una conjunción copulativa o disyuntiva o bien usando otros recursos gramaticales como la especificación, la comparación, la condición, la atribución. Algunos autores [Kister, 1993] denominan *anáfora* a este tipo de abreviación discursiva. Desde otro punto de vista, estas unidades escondidas se podrían considerar un tipo de variación discursiva.

La mayoría de anáforas discursivas de este tipo se producen para agilizar el texto, pues no se repiten los segmentos que corresponden a la parte compartida. En algunas situaciones se elide el núcleo, en otras el complemento o una parte del complemento.

Veamos ejemplos de anáforas extraídas de nuestro corpus sobre enfermedades infecciosas por Rickettsia:

- unidades escondidas en frases o sintagmas coordinados con una conjunción copulativa: *Des del punt de vista clínic, cal fer el diagnòstic diferencial amb* ***malalties víriques i bacterianes.***
- *unidades escondidas en frases o sintagmas coordinados con una conjunción disyuntiva:* **tifus murí** o **endèmic.**
- unidades escondidas en sintagmas especificativos: *Amb certa freqüència (7,5%) se'n presenta formes greus que inclouen diverses combinacions d'****insuficiència orgànica greu (neurològica, respiratòria, renal, cardíaca, hepàtica).***
- unidades escondidas en frases comparativas: *La febre tacada de les Muntanyes Rocalloses respon al tractament*

amb tetraciclines i cloranfenicol tant ***per via oral*** *com* ***intravenosa.***

- unidades escondidas en frases condicionales: *En el 9% de casos sol haver-hi* ***conjuntivitis bilateral.*** *Quan és* ***unilateral*** *i va acompanyada d'afectació ganglionar regional intensa (síndrome oculoglandular) és molt probable que aquesta sigui la porta d'entrada de la infecció.*
- unidades escondidas en frases predicativas: *La febre botonosa és la rickettsiosi exantemàtica més freqüent als països de la conca de la Mediterrània on la* ***malaltia*** *és* ***endèmica.***

2.2 Silencio extrínseco

El silencio extrínseco, causado por la definición misma del objeto del sistema de extracción automática, afecta a un 48% de las unidades que son unidades especializadas de un texto.

Los extractores suelen limitar el objeto de detección a la extracción de la UTP que, por bien que sea la unidad más frecuente de las unidades especializadas, no es la única. La diversidad de las unidades especializadas (por lo que se refiere a su naturaleza, categoría gramatical y estructura) que se usan en los textos especializados conduce a pensar que el objeto de un nuevo concepto de extractor tiene que ser todas las unidades de significación especializada de un texto, y no sólo las unidades terminológicas poliléxicas.

Por eso, hemos analizado todas las unidades especializadas que no eran UTP de un corpus de biomedicina – es decir, USE monoléxicas, siglas, unidades fraseológicas, símbolos, nombres en latín – con el fin de encontrar recursos que ayuden a un extractor a reducir el silencio extrínseco. En esta línea, hemos evidenciado que el dominio de las ciencias de la salud reúne tres características que pueden facilitar el reconocimiento y extracción automatizados de las unidades especializadas de un texto:

1. El uso de nomenclaturas normalizadas. La complejidad de la medicina (en su vertiente teórica, aplicada y práctica) justifica el uso de nomenclaturas médicas relativas a enfermedades, procedimientos diagnósticos, procedimientos terapéuticos. Pero también el uso de nomenclaturas de campos léxicos que constituyen los fundamentos de la medicina: la nomenclatura de la anatomía, de la química, de la zoología, de la botánica, de la bacteriología, de la virología. Las primeras no están tan consensuadas como las segundas; estas últimas, además, están basadas en clasificaciones muy establecidas.
2. El uso de los formantes grecolatinos. El hecho que más del 60% del léxico médico esté basado en un número de raíces, de prefijos y de sufijos grecolatinos limitado avala nuestra propuesta de que un extractor con un diccionario de formantes (la cifra se estima alrededor de los 1.100 constituyentes) con información semántica y reglas de combinación de los formantes reduciría de manera substancial el silencio.
3. El uso de heurísticas de morfología léxica. En concreto, nos referimos al papel que tienen algunos sufijos patrimoniales cuando se adjuntan a unas bases determinadas en el marco de un dominio semántico delimitado. Por citar uno de los ejemplos comentados, los únicos adverbios con significado especializado son los que se han formado a partir de un adjetivo de carácter especializado y el sufijo *-mente* y, normalmente, significan "desde el punto de vista + A".

En resumen, a partir de las conclusiones del análisis, defendemos la idea de que, para diseñar un extractor que sea más exhaustivo, se deberían tener en cuenta todas las unidades de significación especializada que pueden aparecer en un texto especializado, tanto si pertenecen al código de la lengua natural (unidades léxicas y unidades fraseológicas) como si no (símbolos, nomenclaturas), tanto si son simples como complejas, tanto si son léxicas como fraseológicas. Además pensamos que todas es-

tas unidades ofrecen recursos formales (morfológicos y/o sintácticos) y semánticos para su reconocimiento automático.

3. Limitaciones de los extractores: el ruido

En general, los extractores basados en conocimiento lingüístico (sobre todo en patrones morfosintácticos, que son la mayoría) generan unos porcentajes altos de ruido: entre el 45% y el 75% de los candidatos propuestos por estos programas se tienen que rechazar.

Estos resultados obligan a plantearse dos cuestiones: ¿Qué es lo que provoca ruido? ¿Qué tipo de unidades son las que sistemáticamente se "rechazan"?

El factor principal del ruido proviene del concepto mismo de UT con el que los extractores trabajan. En realidad, para estos sistemas el término es una forma exclusiva de un ámbito especializado; y, de su vertiente formal, la mayoría de sistemas sólo tienen en cuenta la estructura sintagmática. Desde el punto de vista lingüístico, pero, una UT es la asociación de una forma y de un contenido y no sólo una forma. Por eso, la estructura formal de una UTP -a pesar de que es un indicio probabilístico–no es un elemento suficiente que sirva para discriminarlas de otras clases de UT.

La causa del ruido reside, pues, en la incapacidad de discriminar las UTP a partir de las estructuras morfosintácticas, y este hecho está suscitado por las restricciones de base de las que parten estos sistemas. Las unidades que comparten las mismas estructuras que las UT son las siguientes:

- UTP (*medul·la òssia, meningitis bacteriana, malaltia de Brill-Zinsser)*
- unidades fraseológicas especializadas *(acumulació de líquid extravascular, augment de la permeabilitat vascular)*
- combinaciones especializadas recurrentes *(radiografia de la mà, massatge a les cervicals, etc.)*
- unidades discursivas *(augment del nombre de casos, dècades dels anys trenta, distribució geogràfica, manera específica, color vermellós, escorxador de Brisbane)*

- UTP no pertinentes para el ámbito temático del texto *(blau de metilè, conca mediterrània, treballadors socials, condicions de vida, classe social, canvi climàtic)*

También puede ocurrir que sólo una parte del segmento sea terminológicamente pertinente, y en ese caso, el carácter pertinente puede tenerlo el núcleo y/o el complemento como muestran los ejemplos siguientes: biòpsia de la pell, presència de rickèttsies, peculiaritats clíniques, existència de la taca negra)

De estas estructuras, las que provocan más ruido, por orden ascendente, son:

- [N [[de art N]$_{SPrep}$ [A]$_{SAdj}$]$_{SPrep}$] SN
- [N [de art N]$_{SPrep}$] SN
- [N [[de N]$_{SPrep}$ [A]$_{SAdj}$]$_{SPrep}$] SN
- [N [de N]$_{SPrep}$] SN

La presencia del artículo delante del complemento es a menudo un indicio de que la unidad no está totalmente lexicalizada y, por tanto, las estructuras en las que el complemento está introducido por un artículo determinante porcentualmente tienden a generar más ruido que las estructuras en las que el complemento es indeterminado. En cambio, las estructuras [N[A]$_{SAdj}$]$_{SN\,y}$ [[N[A]$_{SAdj}$]SN [A]$_{SAdj}$]SN generan menos ruido, lo que no quiere decir que generen poco.

A partir de estas primeras observaciones hemos analizado las clases de adjetivos y las clases de nombres que producen ruido estableciendo las combinaciones pertinentes en el ámbito de la medicina.[3]

[3] Estos análisis, que ahora no podemos exponer, se encuentran especificados en Estopà, 1999 y Estopà, 2001.

4. Elementos para mejorar un extractor

A partir de las constataciones que hemos hecho en relación al ruido y al silencio en el trabajo de tesis doctoral presentado en julio de 1999 en la Universidad Pompeu Fabra, propusimos diversas estrategias para mejorar los resultados de un extractor y hacerlo más adecuado a las necesidades de su utilización (Estopà, 1999; Estopa, Vivaldi, Cabré, 2000).

En esta ocasión remarcaremos solamente dos elementos que nos parecen que mejorarían un extractor: por un lado, el objeto de trabajo y, por el otro, el punto de vista de la pertinencia funcional de este objeto. Estos dos elementos son consecuentes con dos de los supuestos teóricos de la Teoría Comunicativa de la Terminología (Cabré, 1999) en la cual se enmarca este trabajo: a) el hábitat natural de las unidades de significación especializada son los textos especializados y no los diccionarios; b) cualquier aplicación debe adecuarse a las necesidades de utilización.

4.1 El objeto de trabajo

Desde sus orígenes la terminología ha considerado que el término – entendido mayoritariamente como una unidad léxica nominal del lenguaje natural – era su unidad de base. En la última década, pero, algunos cambios de orientación en la terminología teórica y práctica han permitido ensanchar el interés por otras unidades de significación especializada presentes en los textos especializados que van más allá de las UT.

En esta línea, los resultados de los vaciados manuales de un texto de biomedicina realizados por especialistas, refuerzan la idea que la UT no es la única unidad de significación especializada pertinente de los textos especializados. De acuerdo con esta nueva perspectiva, la unidad que es objeto de estudio del extractor que proponemos no puede reducirse ni a la UTP ni a la UT en general, sino que debe abastar todas las *unidades que transmiten significación especializada* (USE), que incluyen tanto las unidades especializadas de categorías gramaticales diferentes que forman

parte del lenguaje natural, como las unidades que forman parte de lenguajes artificiales; y dentro de las unidades que son lenguaje natural, abarca desde las UT simples a las complejas, desde los nombres a los verbos, adjetivos y adverbios, desde las unidades léxicas a las unidades fraseológicas especializadas (UFE); y, finalmente, dentro de las unidades de sistemas artificiales, comprende tanto los símbolos nominales y los nombres latinos propios de nomenclaturas consensuadas como las fórmulas complejas. Estas unidades se recogen en el esquema siguiente:

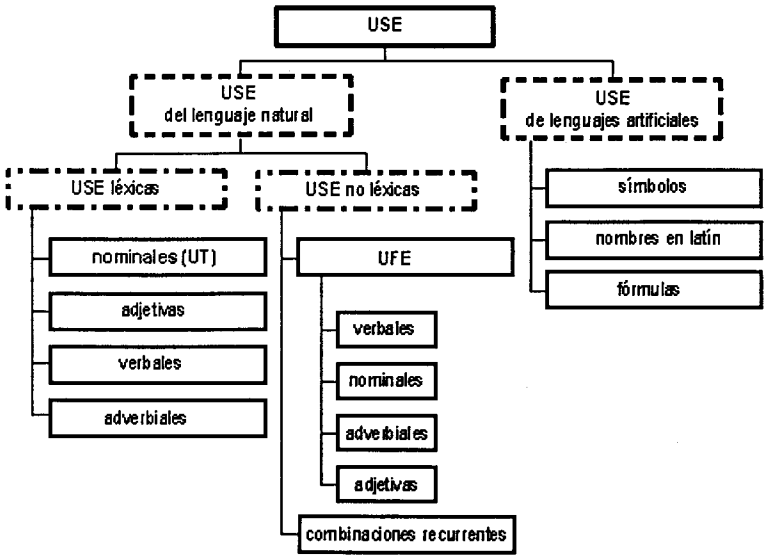

Figura 1

Este cambio de objeto de extracción nos obliga a redefinir los sistemas de extracción automática a candidatos a término clásicos y a orientar el objetivo de nuestro trabajo en la búsqueda y selección de los elementos y estrategias que requiere un sistema de extracción automática de candidatos a unidades de significación especializada (SEACUSE). En este sentido, teniendo en cuenta que el objeto es diverso, las estrategias de detección y extracción también deberían ser diversas y además pensamos que el extractor debería utilizar más de una estrategia para discriminar un tipo concreto de unidad.

4.2 La pertinencia funcional de una unidad

En el apartado anterior hemos puesto en cuestión que la única unidad de interés de un texto especializado fuese la UT y, más concretamente, la UTP; por eso, en contraste con visiones más clásicas y restrictivas, hemos abierto substancialmente el objeto de trabajo.

En este apartado nos preguntamos si las USE son las mismas para todos los usuarios o si contrariamente según las finalidades serán pertinentes determinados tipos de unidades y no otras. Si en el anterior punto nos cuestionábamos el ¿qué? (objeto de estudio) ahora nos planteamos el ¿para quién? (los usuarios de los productos terminológicos) y el ¿para qué? (las necesidades de los usuarios).

Queremos mostrar cómo las USE varían cualitativamente y cuantitativamente en relación con las necesidades profesionales en las que se usan. Partimos, pues, de los supuestos que según los intereses de los diferentes profesionales: no todas las USE que comprende un texto son pertinentes para una determinada actividad profesional. Como consecuencia de estos dos supuestos, las USE de un texto que interesan a un especialista, a un traductor, a un terminógrafo o a un documentalista pueden no coincidir totalmente.

Así, siguiendo con el ejemplo del vaciado automático, un extractor que no tenga en cuenta el punto de vista del usuario - como lo hacen la mayoría de extractores –, aplicado a un texto concreto produce siempre la misma selección de términos. Pero, si se quiere que los sistemas de extracción puedan realizar selecciones adecuadas a las necesidades profesionales de los diferentes colectivos de usuarios, se deben poder perfilar previamente estas necesidades.

Un experimento que hemos realizado (Estopà, 1999) ha mostrado que, ante las distintas posibilidades de selección, cada colectivo profesional elige un conjunto distinto de unidades como pertinentes. Esta selección incluye los tipos de unidades seleccionadas, el tipo más representativo y los parámetros que restringen o priorizan la selección.

La prueba experimental, que se apoyaba en la hipótesis que colectivos profesionales diferentes cuando realizan una determinada actividad profesional se aproximan a los textos especializados con intereses diferentes, pues tienen puntos de vista distintos sobre las USE pertinentes de un texto especializado, consistía en realizar el vaciado de las unidades especializadas pertinentes para diferentes actividades profesionales. En concreto, hemos seleccionado cuatro colectivos profesionales relacionados con los textos especializados: especialistas, documentalistas, traductores especializados y lingüistas/terminógrafos. El vaciado del texto se ha llevado a cabo por tres especialistas de cada uno de los colectivos profesionales propuestos.

De acuerdo con la hipótesis de que es la actividad profesional, y no el colectivo, el elemento pertinente que condiciona la selección de las USE de un texto, hemos restringido las actividades profesionales a una única finalidad diferente para cada colectivo, y no hemos tenido en cuenta que un mismo profesional podía realizar diversas actividades a partir de un texto de especialidad. En este sentido, hemos preferido centrarnos en las actividades profesionales más específicas de cada colectivo: la transmisión del conocimiento (médicos), la indización de textos (documentalistas), la traducción (traductores especializados), y la elaboración de terminologías (terminógrafos).

El texto que hemos utilizado para el vaciado de las USE es el mismo que utilizamos en anteriormente sobre *Enfermedades producidas por Rickettsia.*[4]

Los resultados de esta prueba experimental han demostrado que no todas las USE que hay en un texto son pertinentes para todas las actividades profesionales, sus selecciones sólo coinciden en un 9,3% de las unidades. Esto significa que, primero, las USE (y por tanto también los términos) se modelan en función de las necesidades funcionales de sus destinatarios; y, segundo, la noción de USE pertinente depende de la actividad profesional

[4] Se trata de un texto escrito por especialistas en medicina interna y dirigido a estudiantes de últimos cursos de medicina y a profesionales de la medicina (Farreras-Rozman, 1997).

que se realice, lo que presupone que, desde un punto de vista funcional, las USE de un texto no están prefijadas.

En un texto especializado hay USE pertinentes para todos los colectivos, pertinentes sólo para algunos colectivos, y pertinentes sólo para uno de los colectivos. Los datos globales de los cuatro colectivos profesionales refuerzan la idea de que cada colectivo tiene un criterio propio de selección de USE pertinentes de un texto, y esta diversificación de criterios conlleva, como se puede ver en la tabla, una diversidad en el número de USE seleccionadas y en el tipo de USE priorizadas:

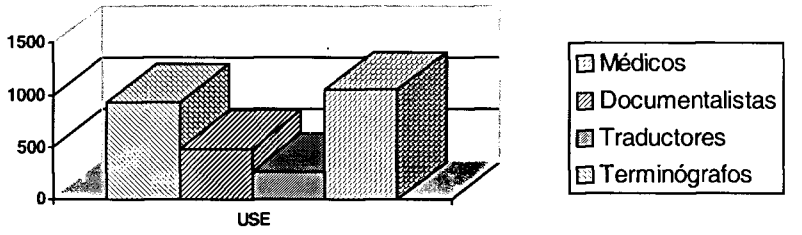

Figura 2: Número total de USE seleccionadas por cada colectivo profesional

En efecto, los resultados obtenidos constatan que el número de USE seleccionadas no coincide entre colectivos. Los médicos han señalado 938, los traductores 270, los documentalistas 486 y los terminógrafos 1.052. Además los colectivos han coincidido muy poco en las unidades que han subrayado. Así, médicos, documentalistas y terminógrafos sólo han coincidido en un 30,7% de unidades, y médicos, traductores y terminógrafos en un 11,3%. Las coincidencias aumentan considerablemente si comparamos los colectivos de dos en dos: médicos y lingüistas coinciden en, médicos y traductores, médicos y documentalistas, lingüistas y documentalistas, documentalistas y traductores.

Esta constatación verifica el supuesto que, desde el punto vista funcional, las USE (y por tanto también los términos) de un ámbito o de un objeto temático no están preestablecidas, sino

que varían en relación con las actividades profesionales. Eso significa que en un mismo texto hay palabras que sólo son pertinentes para un colectivo de acuerdo con la *actividad profesional* que lleva a término.

Otra característica que se desprende de la análisis global de los vaciados es la falta de coincidencia entre colectivos en los tipos de USE marcadas:

	médicos		**documentalistas**		**traductores**		**terminógrafos**	
	Núm.	%	Núm.	%	Núm.	%	Núm.	%
Nombres	824	87,84	426	87,65	211	78,14	900	85,56
Verbos	17	1,81	0	0	1	0,37	49	4,65
Adjetivos	28	2,98	5	1,02	27	10	35	3,32
Adverbios	5	0,53	0	0	2	0,74	4	0,37
Siglas	13	1,38	12	2,46	5	1,85	12	1,13
Símbolos	6	0,63	3	0,61	0	0	6	0,56
Nombres científicos	44	4,69	22	4,52	0	0	45	4,27
Nombres propios	0	0	18	3,70	1	0,37	0	0
Frases discursivas	1	0,10	0	0	23	8,51	1	0,09
Total	**938**	**100**	**486**	**100**	**270**	**100**	**1052**	**100**

Figura 3: tipos de unidades seleccionadas por cada colectivo

En este sentido, constatamos una gama de posibilidades que va desde considerar sólo las USE nominales (como es el caso de algunos documentalistas) a seleccionar todos los tipos de USE presentadas en el capítulo anterior (las selecciones de los médicos). En general, los terminógrafos, seguidos de los especialistas, son los que más unidades han marcado y más variadas, por lo que se refiere a la estructura y a la categoría gramatical. Los documentalistas y los traductores, en cambio, son los que han marcado menos unidades.[5]

[5] Debemos decir que si bien suponíamos que los documentalistas serían los que marcarían menos unidades del texto, pensábamos que los traductores serían los que marcarían más. Pero no habíamos tenido en cuenta la diferencia que el traductor hace entre las USE que contiene un texto y las USE pertinentes para la preparación de su traducción porque presentan problemas de traducción. Y estas últimas son, por suerte, muchas menos que las primeras.

En relación con los aspectos cuantitativos de los vaciados de los médicos, podemos decir que estos han marcado muchas unidades, aunque no siempre de manera sistemática. Parece que en la aplicación de los criterios de selección, tienden a ser muy sistemáticos en la selección de las UT, las USE adverbiales, las siglas, los símbolos, los nombres científicos que forman parte de nomenclaturas consensuadas y las UT con un núcleo deverbal; no lo son tanto, en cambio, en relación a las UFE formadas por combinaciones muy frecuentes de dos unidades terminológicas como tratamiento de + *una enfermedad, radiografía de* + *art* + *parte anatómica*; y son muy poco sistemáticos en relación con las USE adjetivas y verbales, aunque debemos señalar algunas. Finalmente, remarcamos que no marcan UFE verbales ni adverbiales, ni nombres propios aislados.

Al margen de la heterogeneidad en la selección de algunas unidades, los tipos de USE que los especialistas consideran pertinentes, porque son unidades que transmiten conocimiento especializado, son las siguientes: USE nominales, es decir términos, USE verbales, USE adjetivales, USE adverbiales, nombres científicos, símbolos, siglas, UFE nominales, verbales y combinaciones recurrentes nominales.

Por lo que se refiere a los documentalistas, han marcado muy pocas USE del texto en relación con otros informadores y los tipos de unidades seleccionadas se reducen a los siguientes: unidades monoléxicas nominales (42,91%), unidades poliléxicas nominales (55,85%), adjetivos (0,41%), siglas (2,46%), símbolos (0,61%), nombres científicos (pero sólo los nombres de microorganismos) (4,51%), nombres propios (3,69%). No consideran pertinentes ni los adverbios ni los verbos ni las UFE. A diferencia del resto de informadores, dos de los documentalistas han marcado como unidades válidas para indizar un texto nombres propios. Estos nombres se refieren a países en los que se dan las condiciones que provocan una enfermedad determinada, a escuelas de medicina determinadas o a médicos célebres, y facilitan la delimitación precisa de las búsquedas potenciales.

La frecuencia de uso de las USE pertinentes y su ubicación en el texto son dos datos muy relevantes para los documentalistas.

En este sentido, si una USE se da en una frecuencia alta (se va repitiendo en cada párrafo o en cada apartado) y/o integra el título del documento, alguno de sus subtítulos, esquemas o resúmenes, existe una probabilidad muy alta de que se trate de una unidad representativa del texto. Además, los documentalistas tienden a seleccionar las unidades más especificadas (las UTP), porque permiten precisar más y, consiguientemente, reducir el ruido que provocan las unidades muy genéricas o polisémicas.

Así, podemos decir que el vaciado de las USE pertinentes para indizar textos se diferencia de otras finalidades profesionales por: la reducción de categorías gramaticales: sólo nombres y algún adjetivo, el bajo número de USE: sólo las unidades que son representativas del texto (en este sentido la frecuencia y la ubicación de las unidades en el texto son datos imprescindibles), la importancia que pueden tener ciertos nombres propios para identificar un documento, y el predominio de unidades poliléxicas.

Los traductores son el colectivo que ha seleccionado menos unidades. De hecho, sólo han seleccionado las unidades que desconocen semánticamente o que les pueden ocasionar problemas de traducción. Esta es la razón por la que hay una reducción considerable de los tipos de USE seleccionadas: no seleccionan símbolos, no seleccionan nombres científicos en latín, seleccionan muy pocas siglas. Esta restricción es lógica, si tenemos en cuenta que los símbolos y los nombres en latín de las nomenclaturas científicas, en general, son universales y, por tanto, no son objeto de traducción. Las siglas que se usan en medicina aparecen habitualmente en inglés (por su condición de lengua internacional), hecho que elimina cualquier problema de traducción. No obstante, los traductores han marcado algunas de las siglas del texto que, aunque no se traducen, son poco conocidas y las han acompañado del contexto en el que se encuentra su referente: *concentració inhibidora mínima* (CIM).

Otra peculiaridad del vaciado de los traductores es el hecho de que hayan seleccionado las USE en contexto sintáctico porque, muchas veces, es el contexto el que les proporciona elementos lingüísticos y pragmáticos para proponer los equivalentes más adecuados. También han marcado los referentes socioculturales

que incluye el texto, que en la traducción se deberá adaptar o explicar.

Finalmente, en relación con el vaciado, los terminógrafos son los más exhaustivos en la selección de unidades, aunque después no aprovechen todas las USE seleccionadas para elaborar un diccionario concreto.

Es interesante observar que cuando los terminógrafos han marcado verbos se han referido también a su contexto de uso. El contexto puede facilitar ejemplos, en el caso de que el diccionario los incluya, pero también puede indicar la presencia de fraseología verbal. Esta fraseología ha sido tradicionalmente excluida de los diccionarios especializados, pero, en los últimos años, se ha empezado a considerar la pertinencia de introducirla en este tipo de obra.

En definitiva, los datos muestran que las diferencias de los vaciados entre los profesionales son muy significativas tanto en el número de unidades que seleccionan como en los tipos de USE seleccionadas. Esta discrepancia se explica por el hecho de que cada actividad profesional requiere unas USE precisas para realizar sus funciones y a la vez prescinde de otras que pueden ser pertinentes para cualquier otra actividad.

Pero, que la selección de USE sea específica para cada trabajo no presupone que el concepto de USE sea plural. Desde nuestro punto de vista, la noción de USE es única y lo que cambia es el concepto de USE pertinente, de manera que una USE, sin dejarlo de ser, puede no ser pertinente para una actividad concreta. Los análisis de los resultados obtenidos confirman y validan la hipótesis que *la pertinencia de una USE depende de la finalidad profesional.*

4.3 Perfiles de necesidades diferentes

Los análisis cuantitativos y cualitativos de los vaciados manuales de diferentes colectivos con necesidades profesionales concretas nos permiten constituir perfiles que tengan en cuenta por un lado el tipos de unidad y por el otro informaciones sobre estas unidades como la frecuencia de uso, el contexto de uso o la ubicación en el texto.

De acuerdo con estas dos variables estableceremos los cuatro perfiles referentes a las actividades profesionales siguientes: la transmisión del conocimiento especializado (especialistas), la indización de un texto (documentalistas), la traducción especializada (traductores especializados) y la práctica terminográfica (terminógrafos).

Para los especialistas en documentación las USE pertinentes de un texto de especialidad son aquellas que funcionan como identificadores del contenido informativo del texto y que les permiten describir, indizar, ordenar y recuperar un texto especializado determinado. Consiguientemente, sería útil que, cuando un extractor se utilizara para indizar un texto, proporcionara información sobre la frecuencia y sobre la disposición discursiva de las USE en el corpus textual, de manera que sólo mostrase aquellas USE que superasen una frecuencia determinada (como mínimo superior a cinco ocurrencias) acompañadas de información situacional y que respondieran a los tipos siguientes, preferentemente unidades nominales poliléxicas. I además, los nombres propios en contexto de uso.

Las USE que interesan a los traductores son sólo aquellas que les podrían plantear cierta dificultad a la hora de traducirlas: unidades de las que desconocen el significado o unidades que intuyen que ocasionaran problemas de traducción. Por eso, a veces sólo seleccionan segmentos de UTP (y no la unidad entera), sobre todo las unidades (nominales o adjetivas) de carácter no especializado que integren alguna unidad poliléxica. El hecho de que cada traductor tenga necesidades cognitivas, lingüísticas y sociofuncionales diferentes que dependen de su nivel de conocimiento del tema en ambas lenguas conlleva que no haya unos tipos de unidades que interesen más que otros, todo depende de la experiencia profesional.

En este sentido, hemos llegado a la conclusión que las USE que ocasionan más problemas de traducción son: UFE, siglas no internacionalizadas, epónimos, USE o segmentos de USE sin carácter especializado y los neologismos. Por consiguiente, si un extractor debe utilizarse para las necesidades terminológicas de la traducción, es interesante que pueda recuperar las unidades siguientes, todas dentro de su contexto de uso, porque muchas

veces el contexto ofrece datos que facilitan la comprensión de la unidad o la búsqueda de su equivalente.

Para el terminógrafo sería importante que un extractor recuperase todas las USE del texto acompañadas, obligatoriamente u opcionalmente, del contexto y la frecuencia de uso, y relacionase las USE del lenguaje natural de las de lenguajes artificiales.

5. Conclusiones

En conclusión, hemos querido exponer básicamente algunos aspectos tratados en la tesis doctoral Extracción de terminología: *elementos para la construcción de un SEACUSE*. En primer lugar hemos presentado los resultados del análisis crítico de los principales extractores existentes. Después hemos planteado las limitaciones básicas de estos extractores. Y finalmente hemos expuesto dos aspectos nuevos, necesarios para la mejora de un extractor: la abertura del objeto de trabajo y la pertinencia del punto de vista funcional.

El paso de la unidad terminológica a la unidad de significación especializada nos ha permitido considerar como objeto de estudio otras unidades usadas también con valor especializado, categorialmente, sintácticamente y semánticamente diferentes de los términos.

Y la condición de pertinencia de una USE en relación con una actividad profesional nos ha facilitado establecer perfiles profesionales de USE y ha puesto de manifiesto que cualquier aplicación terminológica, para que sea útil, no puede funcionar independientemente de sus finalidades, sino que siempre tiene que adecuarse a las necesidades profesionales de las actividades que realicen sus usuarios.

Estos dos elementos – la abertura del objeto de los textos especializados y la pertinencia o no pertinencia de una unidad – son consecuentes con los principios teóricos de la Teoría Comunicativa de la Terminología en el sentido que permiten adecuar una aplicación a las necesidades de uso real.

6. Agradecimientos

Me gustaría agradecer a la profesora Ieda Maria Alves y a RITERM el haberme invitado a participar en la *I Escuela de Invierno de Terminología*. Esta participación me hace sinceramente mucha ilusión. También agradezco la traducción del resumen a Cleci Bevilacqua.

7. Bibliografía

CABRÉ, M.T. (1999) *La terminologia. Representación y comunicación. Una teoría de base comunicativa*. Barcelona: IULA, Universitat Pompeu Fabra. (Sèrie Monografies, 3).

CABRÉ, M.T.; ESTOPÀ, R. (en prensa) On the units of specialised meaning uses in professional communication. *ITTF, 2001*, 1.

CABRÉ, M.T., ESTOPÀ, R.; VIVALDI, J. (en prensa) Automatic term detection: a review of current systems. In: D. BOURIGAULT, C. JACQUEMIN, M.-C. L'HOMME (eds.) *Recent Advances in Computational Terminology*. Amsterdam, John Benjamins.

ESTOPÀ, R. (1996) *Les unitats terminológiques polilexemàtiques en els lèxics especialitzats: dret i medicina*. Trabajo de investigación de doctorado. Barcelona, Institut Universitari de Lingüística Aplicada, UPF.

_____. (1999). *Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'extracció automàtica de candidats a unitats de significació especialitzada)*. Barcelona, Universitat Pompeu Fabra. Tesis doctoral.

_____. (1999) Eficiencia en la extracción automática de terminología. *Perspectives: Studies in Traductology*, 7, 2, p.277-286.

_____. (en prensa) Los adjetivos en las unidades terminológicas poliléxicas: un análisis morfosintáctico. *Organon*.

ESTOPÀ, R.; VIVALDI, J. (1998) État de la question des systèmes d'extraction automatique de candidats à terme: vers une proposition intégratrice. *Actas de las VII Journées ERLA-GLAT*. Brest, Université de Brest, p.385-410.

ESTOPÀ, R.; VIVALDI, J.; CABRÉ, M.T. (1998) Sistemes d'extracció automàtica de candidats a terme. Estat de la qüestió. *Papers de l'IULA*, Sèrie Informes, 22, p.1-68.

ESTOPÀ, R.; VIVALDI, J; CABRÉ, M.T. (2000) Use of Greek and Latin forms for term detection. In: GAVRILIDOU, G. et al (eds.) *Procedings of Second International Conference on Language Resources and Evaluation.* Atenas, National Technical University of Athens, II, p.855-861.

ESTOPÀ, R.; VIVALDI, J; CABRÉ, M.T. (2000) Extraction of monolexical terminological units: requirement analysis. In: *Second International Conference on Language Resources and Evaluation: Terminology Resources and Computation Proceedings.* Atenas, National Technical University of Athens, II, p.51-56.

ESTOPÀ, R.; VALERO, T. (en prensa) Adquisición de conocimiento especializado y unidades de significación especializada en medicina. *Terminómetro: Problemas terminológicos en medicina.*