

Publicado en:

Perspectives: Studies in Translatology, Volume 7:2, 1999, p. 277-286

Eficencia en la extracción automática de terminología.ⁱ



Rosa Estopà

(Institut Universitari de Lingüística Aplicada) e-mail: rosa.estopa@trad.upf.es

Abstract

This article deals with the degree of efficiency that systems for the automatic extraction of likely terms may have within a particular professional field. Through the use of an empirical test, it analyses whether the expectations of users and, particularly of specialised translators match the results obtained with these systems.

0. Introducción

Un sistema de extracción automática de terminología es un conjunto de programas informáticos que tiene como objetivo extraer automáticamente las unidades terminológicas de un corpus textual informatizado. El resultado de la aplicación de cualquier sistema de extracción automática de términos no es una lista de términos, sino de **candidatos a término**, porque es siempre el usuario, apelando a su competencia cognitiva y a su experiencia pragmática, quien, en último lugar, tiene que acabar de decidir qué candidatos propuestos por el sistema son realmente términos y, además, cuáles són términos pertinentes para su actividad profesional. Así pues, todos los sistemas analizan un corpus textual especializado en soporte electrónico del cual extraen secuencias de palabras, candidatas a término, que el usuario tiene que validar. De manera e la selección **definitiva** de unidades no es automática, sino **asisti**. Podemos concluir que la utilización de la etiqueta *sistema de extracción automática de terminología* no es demasiado precisa. Por eso es preferible hablar de *sistemas de extracción automática de candidatos a término* (SEACAT, en adelante), entendido como un sistema informático que extrae de un corpus textual informatizado un conjunto de secuencias que son **candidatas a ser términos**.

Los SEACAT han tenido desde el primer sistema, del año 1989ⁱⁱ, el objetivo de facilitar una fase del trabajo terminológico: el vaciado. Como otras herramientas de ingeniería lingüística, estos sistemas informáticos están concebidos como una *ayuda* al trabajo terminológico, pues mecanizan algunas tareas del trabajo terminológico que, además, gana en *sistematicidad y rapidez de ejecución*. Sin embargo, al diseñar estos sistemas no se consideraron las necesidades reales de sus diversos usuarios, de manera que, tradicionalmente, los SEACAT se han elaborado y funcionan independientemente de la actividad profesional para la cual se usan.

Un análisis de los principales SEACAT (Estopà, Vivaldi, Cabré, 1997; Estopà Vivaldi, 1998; Cabré, Estopà, Vivaldi, 1999) mostró que, efectivamente, los extractores no tienen en cuenta el objetivo del vaciado ni el marco profesional en el que se aplicará el sistema y extraen de un texto especializado una única lista de unidades. Pero esta única selección no puede adecuarse a las necesidades reales de todos sus usuarios porque, tal y como hemos comprobado a través de una prueba experimental (Estopà, 1999), cada actividad profesional requiere un tipo de unidades de significación especializada específicas.

Además, la selección final de las unidades pertinentes de la lista de candidatos generada por un extractor de terminología es una tarea manual y larga. Los sistemas de extracción automática de terminología generan demasiada cantidad de *ruido* (es decir, unidades generadas por el sistema que no son pertinentes) y, también, demasiada cantidad de *silencio* (es decir, unidades que el programa tendría que haber generado y no lo ha hecho). Además esta característica de los programas de extracción no es sólo desde el punto de vista de las *unidades de significación especializada* (USE) que incluye el texto, sino sobre todo desde el punto de vista de la *pertinencia profesional de las unidades*.

1. Supuestos de partida

En este trabajo partimos de los supuestos que, desde el punto de vista terminológico, en los textos especializados no hay sólo unidades terminológicas *poliléxicas*, sino que también encontramos unidades terminológicas *monoléxicas*. Desde el punto de vista lingüístico, en los textos especializados hay diversos tipos de unidades con significación especializada por su naturaleza, estructura

morfosintáctica y categoría gramatical. Así, las USE abarcan todas las unidades léxicas de los textos especializados usadas con un significado especializado, dentro de las cuales se ubican las unidades terminológicas. Por lo consiguiente, asumimos que los términos sólo corresponden a las USE de categoría gramatical nominal que tienen capacidad de referir y de categorizar la realidad. Y, desde el punto de vista funcional, no todas las unidades de significación especializada que contienen los textos especializados son *pertinentes* para todas las actividades profesionales. Consideramos que la pertinencia de una unidad depende de la actividad profesional, en consecuencia, en un texto especializado no todas las USE temáticamente pertinentes lo son funcionalmente.

En consecuencia, si se quiere que un extractor sea realmente eficaz, veraz y adecuado en su selección, no puede obviar que las unidades terminológicas también pueden ser monoléxicas y que no todas las USE de un texto especializado son válidas para todas las actividades profesionales. Las necesidades especializadas de un traductor especializado no son las mismas, por ejemplo, que las de un documentalista o las de un terminógrafo.

En contraposición con los sistemas tradicionales, defendemos el principio de que un extractor no puede desentenderse de las necesidades de las actividades profesionales para las cuales se utiliza; dicho de otra manera, no puede omitir la pertinencia o no pertinencia de las USE para una actividad concreta.

2. Objetivo

El objetivo principal de este estudio es mostrar las insuficiencias de un SEACAT tradicional, basado en patrones morfosintácticos que no tiene en cuenta la finalidad profesional del vaciado, lo que es característica mayoritaria. El estudio adopta una perspectiva profesional concreta: la de la traducción especializada. En concreto queremos comprobar si los sistemas de extracción automática de terminología responden a las necesidades de los traductores especializados.

Para llevarlo a cabo nos planteamos las tres cuestiones siguientes: ¿Qué tipos de unidades detectan los SEACAT? ¿Qué necesidades tienen los traductores especializados desde el punto de vista de la terminología cuando preparan una traducción? ¿Cuáles son las deficiencias de los SEACAT desde la perspectiva de las necesidades del traductor especializado?

3. Unidades detectadas por los SEACAT

Para saber qué tipos de unidades detectan sistemáticamente los SEACAT nos hemos basado, por un lado, en el estudio del funcionamiento general de dieciocho extractores existentes (Estopà, Vivaldi, Cabré, 1997; Estopà y Vivaldi, 1998; Cabré, Estopà y Vivaldi, 1999) y, por el otro, en el análisis de las unidades resultantes de la aplicación de un extractor de base lingüística (Estopà, 1999).

En aquel estudio de los principales SEACAT existentes constatamos que:

1. El objeto de base de la mayoría de SEACAT es la unidad terminológica poliléxica (UTP) exclusivamente.
2. Todos los SEACAT analizados generan demasiada cantidad de *silencio* y de *ruido*.
3. Los SEACAT producen dos tipos de silencio diferente: silencio *intrínseco* al objeto de vaciado del sistema de extracción (no detectan las unidades anaforizadas discursivamente); y silencio *extrínseco* al objeto de vaciado del sistema de extracciónⁱⁱⁱ (sólo detectan unos tipos muy limitados de unidades de significación especializada).
4. La mayoría de los SEACAT se basan exclusivamente en la forma del término.
5. La mayoría de los SEACAT se basan en patrones morfosintácticos planos para identificar los términos complejos puesto que no usan analizadores sintácticos.
6. La intervención del usuario al final del proceso es imprescindible para seleccionar las unidades candidatas.

Por otro lado, el análisis de un corpus de medicina de 70.000 ocurrencias nos permite afirmar que los extractores detectan diversas clases de segmentos que responden a uno de los patrones prototípicos de las unidades terminológicas poliléxicas. Ahora bien, aunque todos los segmentos detectados corresponden, efectivamente, a un patrón morfosintáctico esperado, como la estructura de las unidades terminológicas poliléxicas no es exclusiva de estas unidades, también se detectan otros tipos de unidades no terminológicas. Así, una unidad

seleccionada por un extractor puede responder a cualquiera de los siguientes tipos:

- Unidades terminológicas poliléxicas (UTP) (*medula ósea, meningitis bacteriana, enfermedad de Brill-Zinsser, prueba de la immunoperoxidasa, sistema mononuclear fagocítico, etc.*).
- Unidades fraseológicas especializadas (UFE) (*acumulación de líquido extravascular, aumento de la permeabilidad vascular, extravasación de líquido intravascular, factor de necrosis tumoral, etc.*).
- Combinaciones especializadas recurrentes (*radiografía de la mano, masaje en las cervicales, etc.*).
- Unidades discursivas (UD) (*aumento del número de casos, décadas de los años treinta, distribución geográfica, manera específica, resultados retardados, baños frecuentes, tiempo adicional, estado general, matadero de Brisbane, etc.*).
- Unidades de significación especializada (USE) no pertinentes para el ámbito temático del texto (*azul de metilé, cuenca mediterránea, trabajadores sociales, condiciones de vida, clase social, cambio climático, etc.*).

E, incluso, puede pasar que sólo una parte del segmento sea especialmente pertinente (*biopsia de la piel^{iv}, presencia de rickettsias, peculiaridades clínicas, existencia de la mancha negra, zoonosis de distribución mundial, duración del exantema, tratamiento del tifus epidémico, prevención de la enfermedad, importancia epidemiológica, tratamiento antimicrobiano adecuado, etc.*). En estos casos hablamos de una mala delimitación de la unidad.

4. ¿Qué necesidades tienen los traductores especializados desde el punto de vista terminológico cuando preparan una traducción?

Podemos afirmar de manera reducida que la preparación de una traducción está dividida en dos fases. Una primera fase de comprensión del texto de partida en la que el traductor se encuentra con problemas de cognición y una segunda de búsqueda de equivalentes. Teniendo en cuenta este proceso podemos decir que el texto presenta al traductor problemas lingüísticos, sociolingüísticos y extralingüísticos, entre otros.

En el marco de la traducción especializada y desde la óptica de la segunda fase del proceso de traducción, las USE pertinentes son *unidades de equivalencia*; en general, se trata de las unidades que permiten al traductor traspasar adecuadamente una idea especializada de una lengua a otra. Los traductores ponen, pues, un énfasis especial en la manera de expresar una determinada idea para que la traducción sea correcta y adecuada.

Pragmáticamente, no puede elaborar una terminología sináptica cada vez que un traductor aborda una traducción especializada, al menos por dos razones: en primer lugar, porque no sería efectivo pues, profesionalmente, siempre dispone de un tiempo bastante limitado para hacer la traducción; en segundo lugar, porque sería redundante en relación con sus trabajos anteriores. En este sentido, cuanta más experiencia tenga un traductor en un dominio determinado, menos voluminoso será el vaciado de unidades de significación especializada antes de cada nueva traducción; y contrariamente, cuanto más novedoso sea un tema para el traductor más extenso tendrá que ser este primer vaciado.

Por lo tanto, las USE que interesan al traductor especializado son sólo las que les podrían plantear cierta dificultad a la hora de traducirlas: unidades de las cuales desconoce su significado o unidades que intuye que le ocasionarán problemas lingüísticos o sociolingüísticos de traducción. Por eso, muchas veces sólo seleccionan segmentos de una UTP y no la unidad entera, sobre todo cuando se refiere a unidades nominales o adjetivas de carácter no especializado que integran una unidad especializada más compleja.

Pero, el hecho de que cada traductor tenga necesidades cognitivas, lingüísticas y sociolingüísticas diferentes que dependen, esencialmente, de su nivel de conocimiento del tema en ambas lenguas de traducción, conlleva que no haya unos tipos de unidades que interesen más que otros; todo depende de su experiencia profesional y, para la extracción automática, este criterio es demasiado subjetivo.

A pesar del alto grado de subjetividad del vaciado en la preparación de una traducción pensamos que se puede establecer un perfil general del tipo de unidades que de un texto especializado interesan al traductor especializado. Para comprobarlo, hemos realizado la siguiente prueba experimental: hemos dado un mismo texto sobre las enfermedades infecciosas a dos traductores especializados en medicina para que vaciasen las unidades de significación especializada.

Además estos traductores son profesores de traducción científica en una facultad de traducción.

El texto que hemos utilizado para esta prueba es un documento de medicina interna de 12.069 ocurrencias (*Malalties produïdes per Rickettsia*), extraído de un manual especializado (*Medicina interna*, Farreras-Rozman, 1997).

Para realizar el vaciado de la manera más neutra posible, se les formuló la siguiente consigna: “marcar todas las unidades especializadas del texto que marcarías antes de abordar su traducción.”

Los resultados cuantitativos de los vaciados manuales de los dos traductores se reflejan en la tabla siguiente:

	<i>Traductor 1</i>	<i>Traductor 2</i>
Nombres	137	105
Verbos	1	0
Adjetivos	25	2
Adverbios	2	0
Siglas	2	4
Nombres en latín	0	0
Símbolos	0	0
Sintagmas o frases	15	9
Total	182	120

Analizando los tipos de unidades seleccionadas por los traductores y comparándolos, por ejemplo, con los de la selección que del mismo texto hacen los especialistas o los documentalistas (Estopà, 1999; Cabré, Estopà, 1999) llegamos a las siguientes conclusiones: a) Hay USE que no interesan a un traductor porque no se traducen; básicamente, las unidades de significación especializada no lingüísticas, como por ejemplo, símbolos, nombres de nomenclaturas estandarizadas, siglas muy internacionalizadas. b) Todos los elementos textuales que faciliten la búsqueda de un equivalente de una unidad de significación especializada son pertinentes para un traductor. Muchas veces el contexto de una unidad permite solucionar el problema (cognitivo, lingüístico o sociolingüístico). Los marcadores textuales de cualquier tipo ofrecen datos que facilitan la comprensión de la unidad o la búsqueda de su equivalente. Por consiguiente, si un extractor tiene que servir a las necesidades especializadas de la traducción, es interesante que pueda recuperar las unidades pertinentes dentro de su contexto de uso.

Del análisis de las unidades seleccionadas también hemos llegado a la conclusión de que las unidades de significación especializada que plantean más problemas de traducción y que, por tanto, a menudo, forman parte del vaciado del traductor, son: unidades fraseológicas especializadas; combinaciones especializadas recurrentes; siglas no internacionalizadas; epónimos; segmentos de unidades de significación especializada sin carácter especializado; neologismos que pueden responder a cualquier tipo de unidad lingüística.

5. ¿Cuáles son las deficiencias de los SEACAT desde la perspectiva de las necesidades del traductor especializado?

Los extractores actuales, para la traducción especializada, generan cierta cantidad de ruido, pero sobre todo generan **mucho silencio** extrínseco al objeto de extracción básico de estos sistemas, pues sólo suelen detectar USE poliléxicas, es decir unidades terminológicas poliléxicas, unidades fraseológicas especializadas nominales y combinaciones especializadas nominales recurrentes. No existe, pues, correspondencia entre el material que proporcionan los SEACAT y lo que necesitan obtener de los textos especializados los traductores porque tienen interés en tipos de unidades especializadas mucho más amplias que las que detectan actualmente la mayoría de extractores.

Además, normalmente los extractores suelen presentar los candidatos a término aisladamente, sin ningún tipo de información complementaria que facilite su selección. A los traductores les interesa, por un lado, que las unidades estén contextualizadas; y, por otro, tener acceso a la red de unidades morfológicamente relacionadas del texto. Estas informaciones les ayudan a decidir si es pertinente o no una unidad seleccionada y les facilitan la búsqueda del equivalente más adecuado.

Así pues, un SEACAT que quisiera satisfacer las expectativas del traductor especializado tendría que

1. Extraer de un texto las USE nominales (es decir los términos), pero también las USE verbales, adjetivas y adverbiales y las unidades fraseológicas especializadas (UFE).

2. Extraerlas acompañadas de su contexto de uso, el cual puede facilitar en algunos casos (a través de una definición, una paráfrasis, el artículo, la glosa de una sigla) la solución del problema de traducción.
3. Establecer relaciones conceptuales entre las unidades de significación especializada del texto, para poder detectar y controlar las posibles variantes.
4. Relacionar las unidades del texto que pertenecen a una misma familia morfológica, para poder tomar decisiones sobre un posible equivalente.
5. Clasificar las unidades de significación especializada poliléxicas según el carácter de su núcleo y de su complemento, porque, como es sabido, las unidades que conllevan más problemas de traducción son aquellas en las que uno de sus componentes no tiene carácter especializado por sí mismo, de manera autónoma.

Como opciones complementarias, pensamos que un extractor útil para el traductor especializado también debería:

6. Estar conectado a bases de datos y diccionarios electrónicos, para facilitarle la tarea de toma de decisiones sobre una determinada unidad.
7. Permitir almacenar las unidades problemáticas para elaborar un *glosario del traductor*.
8. Tener una memoria individual de usuario, de manera que, cada vez que el traductor utilice el sistema para preparar una traducción, detecte las unidades ya tratadas anteriormente y las presente con la remisión al glosario del traductor.


6. Conclusiones

En este artículo hemos abordado la cuestión de la eficiencia de los *sistemas extracción automática de candidatos a término* (SEACAT) en el marco de una actividad profesional específica. Hemos planteado si las expectativas de usuarios concretos, en este caso, de traductores especializados, ante un SEACAT clásico coincidían con los resultados reales que generan estos sistemas.

Primero hemos comprobado que los SEACAT actuales no son satisfactorios principalmente porque:

- a) Generan demasiada cantidad de *ruido* pues las estructuras morfosintácticas de las unidades de significación especializada (USE) nominales poliléxicas en las que se basan la mayoría de sistemas no son exclusivas de este tipo de unidades.

- b) Generan demasiada cantidad de *silencio intrínseco* al objeto de extracción porque no tienen mecanismos para detectar las USE poliléxicas discursivamente anaforizadas.
- c) Generan *silencio extrínseco* porque sólo detectan las USE nominales poliléxicas y no las monoléxicas.

En cuanto a la cuación de los SEACAT a las necesidades profesionales, hemos visto que a pesar de que el objetivo de los SEACAT actuales es facilitar el trabajo terminológico al profesional, la mayoría de sistemas no tiene en cuenta que la finalidad profesional condiciona la selección de unidades especializadas de un texto.

Por consiguiente, hemos analizado los desajustes entre lo que detectan los SEACAT y lo que realmente querría que detectasen un colectivo profesional concreto con unas necesidades profesionales específicas: los traductores especializados. En este sentido, hemos visto cómo, desde el punto de vista de la finalidad traductora, los extractores generan unidades innecesarias y, en cambio, silencian muchas otras unidades imprescindibles porque sus necesidades especializadas van mucho más allá de las unidades terminológicas propiamente dichas, así abarcan las USE adjetivas, verbales y adverbiales y las fraseológicas.

En conclusión, desde la perspectiva de las necesidades profesionales del usuario, pensamos que un extractor no se puede limitar a generar una lista de palabras, sea cual sea su finalidad de uso, sino que debe adecuar al máximo sus selecciones al perfil de la actividad en el contexto en el cual va a ser usado. Sólo de esta manera podríamos afirmar que realmente los extractores facilitan el trabajo terminológico al profesional, objetivo principal de estos sistemas.

7. Bibliografía

CABRÉ, M. T. (1992): *La terminologia. La teoria, els mètodes, les aplicacions*. Barcelona. Empúries.

CABRÉ, M. T. (1998) “Elementos para una teoría de la terminología: hacia un paradigma alternativo”. *El llenguatge*, 1, 1, 59-78.

ESTOPÀ, R.; (1999) *Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a unitats de*

Significació Especialitzada). Tesis doctoral. Barcelona: Institut universitari de Lingüística Aplicada.

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. Pàgs. 1-68.

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, 385-410.

PERRON, J. (1989) "Termino: un système de dépouillement terminologique". *Terminogramme*, 54, 3-9.

ⁱ Este trabajo es fruto de la tesis doctoral *Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada)* dirigida por la Dra. M. Teresa Cabré que presenté en el Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra en julio de 1999. Es un artículo hecho en el marco del proyecto de investigación *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica* (PB-96-0293). Quiero dar las gracias a Andreína Adelstein y a Judit Feliu por sus lecturas atentas durante el caluroso mes de agosto, y a Cristina Corcoll por la traducción del resumen.

ⁱⁱEl primer prototipo de un SEACAT —TERMINO— se elaboró el año 1989. TERMINO se concibió en el Centre ATO (Analyse de Textes par Ordinateur) de la Universidad de Quebec en Montreal dirigido por Pierre Plante. Los autores lo definían como un "*système de dépouillement assisté par ordinateur*" [Perron, 1989: 6]. Anteriormente, otros proyectos aplicados al campo de la indización realizaban trabajos parecidos [Pratt, A.; Pacak, M., 1969].

ⁱⁱⁱ Entendemos por *silencio intrínseco al sistema* el conjunto de segmentos especializados que no detecta un SEACAT y que tendría que detectar en relación con su objeto de detección; es decir, aquellas unidades terminológicas poliléxicas (UTP) que aparecen en el texto, que son objeto de detección, pero que no las reconoce y, por lo tanto, decimos que las silencia. Y entendemos por *silencio extrínseco al sistema* el conjunto de unidades especializadas que un SEACAT ignora explícitamente; es decir, unidades que, aunque semánticamente son especializadas y aparecen en el texto, formalmente no forman parte de sus objetivos de extracción, a pesar de que el usuario las considera unidades especializadamente pertinentes.

^{iv} Hemos marcado en negrita las unidades con significación especializada.