UNIVERSITAT JAUME·I

Jornades de Foment de la Investigació

ANÁLISIS DE LA EXTRACCIÓN AUTOMÁTICA DE TÉRMINOS CON EL PROGRAMA INFORMÁTICO EXTRATERM

Autors

Paloma BENAVENT
Sara PARRILLA.

1. RESUMEN

A partir del corpus electrónico de textos de especialidad del ámbito de la cerámica, TXTCeram, se ha puesto a prueba la eficacia de ExtraTerm, la herramienta informática de extracción terminológica del paquete de TRADOS 5.5. En el presente artículo se describen los pasos que hemos llevado a cabo para comprobar las posibilidades que esta herramienta nos proporciona a la hora de extraer automáticamente terminología especializada de un corpus electrónico. Asimismo, se exponen los resultados obtenidos de la extracción a partir de un análisis basado en los criterios léxico-semánticos de selección de términos desarrollados por L'Homme (2005); además de valorar tanto las ventajas e inconvenientes como la eficacia de esta herramienta y su posible utilidad para los traductores y terminólogos.
En nuestro caso, los resultados de la extracción de términos e información conceptual contribuirán a la creación de la base de datos de conocimiento OntoCeram, en el marco del proyecto ONTODIC, en el que además se pretende diseñar una metodología de trabajo que sea útil al terminólogo y también al traductor.

2. INTRODUCCIÓN

Durante los últimos años se han desarrollado sistemas de extracción automática de terminología con el fin de facilitar el trabajo de vaciado terminológico a los profesionales de esta disciplina. Estos sistemas pueden definirse como "un conjunto de programas informáticos que reconoce y extrae las unidades terminológicas (UT) que aparecen en un corpus de textos especializado" (Estopà, 2000).
Aunque los sistemas de extracción automática agilizan el proceso de vaciado de términos, se han constatado, en diferentes estudios, los principales defectos con los que cuentan estos programas. Por un parte, generan demasiada cantidad, tanto de ruido (aquellas unidades que no deberían aparecer en el listado de candidatos a término), como de silencio (términos que no han sido recogidos por el programa); y por otra parte, existe cierta dificultad a la hora de detectar las unidades monoléxicas, ya que estos sistemas se centran, sobre todo, en la identificación de unidades poliléxicas, dada la dificultad de discernir los términos simples de las unidades que pertenecen al léxico general. Como consecuencia, la figura del terminólogo aún sigue siendo muy importante, ya que se hace necesaria su participación a la hora de validar el listado de candidatos obtenidos por el extractor terminológico.

Los sistemas de extracción automática se basan en distintas estrategias metodológicas para realizar la extracción de terminología. Las dos estrategias principales son la lingüística y la estadística, aunque también se ha empezado a trabajar con la híbrida, que consiste en una combinación de las ya citadas con el fin de superar las limitaciones con las que éstas cuentan. A continuación se comentan las estrategias metodológicas y se nombran algunos programas basados en ellas (Vivaldi y Rodríguez 2001):

lingüística: la extracción se basa, sobre todo, en patrones sintácticos (TERMS, TERMINO, FASTR o LEXTER)

estadística: bajo este enfoque se enmarcan programas que abarcan desde aquellos programas que se centran en un recuento de frecuencia de uso de palabras hasta aquéllos en que combinan estadísticas que miden, por ejemplo, la colocación de las unidades (ANA).

híbrida: en ésta se intentan integrar las anteriores estrategias, en mayor o menor medida, tomando las características que más interesen en cada caso (ACABIT o TRUCKS).

3. NUESTRO TRABAJO

En este apartado vamos a explicar los diferentes recursos con los que hemos contado para realizar nuestra investigación, la metodología utilizada y el análisis de los resultados obtenido gracias a la extracción realizada con ExtraTerm.

3.1. Recursos utilizados

En nuestro caso decidimos trabajar con la herramienta ExtraTerm, perteneciente al paquete de Trados, versión 5.5, con el fin de que nos extrajera los candidatos a término que se encuentran en el corpus electrónico TXTCeram y, de esta forma, comprobar la eficacia de dicha herramienta. A continuación comentamos brevemente las características tanto del corpus TXTCeram como de la herramienta de extracción automática ExtraTerm.

3.1.1. Corpus electrónico de textos del ámbito de la cerámica industrial TXTCeram

El proyecto TXTCeram contemplaba la creación del corpus electrónico de textos de especialidad del ámbito de la cerámica. El corpus consta de una selección de 25 obras especializadas (entre libros, manuales, revistas monográficas, así como algunos folletos de carácter divulgativo o comercial) del dominio de la cerámica industrial en español. El corpus contiene casi 2,5 millones de palabras (exactamente 2.340.161 palabras), distribuidas en 114 fichero, que están disponibles tanto en formato de texto plano (formato .txt), como en formato de Microsoft Word (formato .doc).

El proceso de digitalización de las obras para compilar el corpus lingüístico se llevó a cabo gracias a la colaboración de estudiantes de las asignaturas de Informática aplicada a la traducción y Terminología, impartidas dentro de la Licenciatura en Traducción e Interpretación de la Universitat Jaume I, así como de estudiantes de doctorado y becarios del Proyecto CREC.

3.1.2. Herramienta de extracción automática ExtraTerm

La extracción automática de terminología se ha realizado con el programa ExtraTerm de Trados, versión 5.5. Esta herramienta permite extraer candidatos a término, presentarlos en una lista y exportar los datos para incorporar los candidatos a término extraídos a las bases de datos terminológicas de MultiTerm 5.5.

La herramienta para realizar la extracción automática se basa, sobre todo, en un análisis estadístico. Este análisis es meramente formal, es decir, se realiza una búsqueda de aquellas unidades léxicas que aparecen en

el corpus con una frecuencia determinada; y además, este proceso se complementa con una serie de técnicas probabilísticas.

3.2. Metodología de trabajo

A lo largo de este apartado explicaremos los pasos que se han seguido a partir de la preparación de ExtraTerm y con el fin de obtener candidatos a término de forma automática.

3.2.1. Preparación de la extracción con ExtraTerm

Para realizar la extracción automática con ExtraTerm es necesario crear un nuevo proyecto en el que se han de configurar una serie de parametrizaciones estándares, como el idioma de los documentos, con el fin de realizar extracciones monolingües o multilingües, o la ubicación de los archivos que se generan durante la extracción. A continuación, se seleccionan los archivos con los que se realizará la extracción y se especifican el resto de las parámetros de configuración: la longitud y el número de candidatos a término a extraer y las listas de palabras vacías y de exclusión a tener en cuenta durante el proceso. Las listas de palabras vacías y de exclusión son listados de palabras que contienen unidades léxicas que no se desean incluir en el proceso de extracción.

Partiendo de la configuración que acabamos de comentar, nuestro proyecto de extracción automática se ha diseñado de la siguiente forma:

- Documentos utilizados: 114 archivos de texto
- Idioma del proyecto: monolingüe (español);
- Longitud máxima de los candidatos a término: 10 unidades;
- Número de candidatos a término a extraer: sin límite;
- Margen de extracción: puntuación entre 50% y 100%;
- Listas de palabras vacías y de exclusión creadas a partir de la lista *stopwords_es.txt* del programa, que contiene 272 palabras, entre artículos, preposiciones, conjunciones, pronombres, además de algunos adjetivos y adverbios de uso frecuente.

3.2.2. Realización de la extracción, exportación y conversión de los resultados

Tras la configuración del proyecto, se procedió a la extracción automática de los candidatos a término. Como resultado de dicho proceso, se obtuvieron 63.474 candidatos a término de las 2.340.161 palabras que contiene el corpus, lo que supone un 2,71%. Todos estos candidatos a término se recogieron en seis archivos que contenían, cada uno de ellos, unos 10.000 candidatos ordenados alfabéticamente.

A continuación y con el fin de validar los candidatos a término, los resultados de la extracción automática se exportaron a una base de datos terminológica de MultiTerm 5, en la que se creó una entrada terminológica para cada candidato a término importado. Como la validación de los candidatos a términos se iba a realizar con la versión SDL MultiTerm 7, posteriormente fue necesario convertir la base de datos terminológica a esta versión del programa, para lo que se utilizó el programa MultiTerm Convert.

3.2.4. Validación de términos

Como acabamos de comentar, la extracción automática nos dio como resultado unos 63.474 candidatos a término, de los cuales hemos analizado 1.850. Esta muestra de candidatos a término procede del estudio de dos de los archivos anteriormente mencionados, de uno se analizaron los primeros 200 candidatos a término, mientras que del otro, se escogieron 1.650, sin tener en cuenta la puntuación de los mismos, que podía variar entre 50% y 100%.

Para analizar los resultados obtenidos por ExtraTerm y, de esta forma, validar los términos proporcionados, hemos utilizado el programa de gestión de bases de datos terminológicas MultiTerm 7 de SDL Internacional. En la imagen se muestra la estructura que presenta una ficha terminológica en Multiterm 7:

Con el fin de validar los candidatos a término, hemos aplicado los criterios léxico-semánticos de selección de términos establecidos por L'Homme (2005):

a. Entidad del ámbito de la cerámica: la unidad extraída designa una entidad (material, programa, entidad de representación, unidad de medida o un ser animado) del ámbito de la cerámica (ej. *cemento, baldosa, decoración, granulometría*)

b. Unidad predicativa que reenvía a una entidad del ámbito de la cerámica: si se trata de unidades predicativas –verbos, nominalizaciones, adjetivos, etc.– se considerarán válidos si los agentes reenvían a entidades de criterio A (ej. cocer: el *bizcocho* se cuece en un *horno de cocción rápida*; material: material *sellante*). Sin embargo, la misma unidad predicativa se puede combinar con agentes no especializados; si esta tiene el mismo sentido que los otros agentes, no se considera válida.

c. Derivado morfológico de un término: si se trata de derivados morfológicos se considerarán válidos si están semánticamente emparentados con un término seleccionado en función de los criterios A o B (ej. *sellar: sellado, sellante, sellador; cocción: cocer, bicocción, cocido*).

d. Unidad que establece una relación paradigmática con otro término: si se trata de una unidad léxica que tiene una relación paradigmática (que no sea una relación morfológica ya identificada en C) con un término seleccionado en función de los criterios A, B o C, se considerará válido. Por ejemplo, si el adjetivo extrusionado se considera válido porque se combina con agentes especializados (aplicación del criterio B), *extrudido* deberá igualmente considerarse válido porque se opone a extrusionado en algunas de sus acepciones (ej. *gres extrusionado; gres extrudido*).

3.3. Análisis de los resultados obtenidos

Una vez aplicados los criterios anteriormente mencionados, se han validado un total de 197 términos de los 1.850 candidatos analizados, lo que resulta ser un porcentaje muy bajo de fiabilidad (sólo un 10,64%). Por una parte hemos obtenido candidatos a término que no deberían haber aparecido, lo que se denomina *ruido*;

mientras que, cuando hemos analizado los contextos de los candidatos a término para la validación de los mismos, se ha hecho patente el *silencio* generado por ExtraTerm, es decir, términos que deberían de haber sido recogidos por el programa y que, por causas desconocidas, no se han extraído.
En primer lugar, vamos a comentar una serie de ejemplo sobre el *silencio* generado en la extracción:

• ***aburjado***: criterio A, entidad del ámbito de la cerámica, y criterio D, unidad que establece una relación paradigmática con otro término (se contrapone a *arenado*)
• ***arena de la clase A, arena de la clase B, arena de la clase C***: criterio D, unidades que establecen una relación paradigmática con otro término (*arena*)
• ***arrastrado***: criterio A, entidad del ámbito de la cerámica (alisado de márgenes exteriores de los objetos secos)

Aún no hemos podido calcular el *silencio* generado por el sistema de extracción automática, ya no se trata del objetivo principal de este estudio, y consideramos que se deberían realizar estudios comparativos para poder hacer un cálculo lo más exacto posible.
Asimismo, es destacable la aparición de lo que podríamos denominar "unidades ambiguas", es decir, unidades que, en algunos contextos, el concepto pertenece al léxico general y que, a su vez, en otros contextos, designa un concepto del campo de cerámica, por lo que se puede validar como término, teniendo tan sólo en cuenta los contextos pertinentes (ej. *arrancado*: defecto del producto intermedio); aunque este tipo de unidades son difíciles de detectar para los programas de extracción automática porque no se trata de una distinción morfológica, sino conceptual.
En cuanto al *ruido* generado por el sistema de extracción automática, hemos clasificado las unidades en varios grupos que se comentan a continuación:

1. Unidades no lingüísticas. Candidatos a términos que no tienen un valor lingüístico y, a su vez, tampoco designan ningún concepto. (Ej. *1:1 o 1:2, ABAB, ABAC, ABO, ADR, AIIb, AIII M*).

2. Errores ortotipográficos. Se trata de candidatos a término que contienen tanto errores ortográficos (Ej. *abrasion, absorvido, argon*), como tipográficos (Ej. *act?, actuaci?, adem?, admisi?, adici?, abr*) debido a errores producidos en la fase de digitalización.

3. Léxico general. Muchos de los candidatos a términos no se han validado, ya que no se consideran términos propios de ninguna especialidad, pues forman parte del léxico general. (Ej. *abuelo, abril, abeja, arriba, abandonar, armario, adaptar*).

4. Términos especializados de otros ámbitos diferentes a la cerámica. Términos que pertenecen a otras disciplinas relacionadas con la cerámica, como la óptica (*aberración, aberración cromática, abertura del diafragma, abertura numérica*), la física (*aceleración centrípeta*) o la arquitectura (*arquitectura modernista,*

arquitectura popular).

5. Candidatos a término en otros idiomas. A pesar de que se trata de un corpus lingüístico en español, hemos encontrado una serie de candidatos a término que aparecen en inglés. (Ej. a*bsorption, act, air, advanced, affecting*).

6. Formas flexionadas. A pesar de que el programa discrimina las variaciones gramaticales de una misma palabra, el programa ha recuperado algunas unidades léxicas sin lematización.

· **Sustantivos** argamasa, argamasas; accesorio, accesorios; acuosa, acuosas.
· **Verbos** *acabar: acabe, acabó, acaban, acabamos, acabarse; añadir: añada, añade, añaden, añadía, añadido, añadiendo, añadiéndole, añadiéndose, añadieron, añadimos, añadió, añadiría, añadirá, añadirla, añadirse; abandonar: abandonada, abandonado, abandonan.*
· **Adjetivos** a*decuado, adecuados, adecuadas, adaptada, adaptadas.*

7. Unidades lingüísticas sin concepto. En muchos casos se proponen como candidatos a término cadenas de palabras con una alta puntuación, pero que no constituyen términos (Ej. *arquitecto debe especificar las juntas de dilatación e indicar, absoluta falta de afinidad entre la heterogeneidad y la fase, actuaciones especiales en zonas que no quedaron limpias la primera*).

8. Unidad perteneciente a un término complejo. Aparecen unidades que por sí solas no designan ningún concepto del ámbito de la cerámica, pero junto con otras unidades forman parte de un término complejo (Ej. arábiga goma arábiga, arena polimíctica arena polimíctica arcillosa)

4. CONCLUSIONES

Basándonos en nuestro estudio y teniendo en cuenta que la muestra analizada es aleatoria, ordenada alfabéticamente y se encuentra entre intervalo de 50% y 100% de puntuación, hemos llegado a las siguientes conclusiones:
- Se ha producido un exceso de ruido debido a que el ExtraTerm, sobre todo, no ha reconocido las formas flexionadas de los verbos y sustantivos como una sola forma y las ha incluido como diferentes candidatos a término.
- A su vez, se puede apreciar silencio en el resultado de la extracción, ya que hemos podido comprobar en los contextos la aparición de posibles términos que, debido a su escasa aparición en el corpus, no han sido recogidos a la hora de proponerlos como candidatos a término.
- Cuando se configuró el programa para la extracción de los candidatos a término, se indicó que éstos contaran con un máximo de 10 unidades, lo que ha llevado a que aparezcan cadenas de palabras muy largas sin representación conceptual.

- Se ha demostrado que el intervalo de extracción de candidatos a término (de 50% a 100% de puntuación) es demasiado amplio.

Para intentar que la herramienta ExtraTerm se muestre más útil y eficaz al extraer terminología específica en el ámbito de la cerámica, proponemos una serie de mejoras, sobre todo a la hora de configurar el programa para la extracción de los candidatos:

1. Elaboración de una lista de palabras vacía más estricta y ficheros de lemas más precisos, en la que se incorporen las palabras gramaticales que se han detectado a lo largo del proceso de validación, y, de esta forma, conseguir un listado más depurado de candidatos a términos.
Con la elaboración de una stoplist más amplia que incluya los verbos vacíos de significado léxico, las abreviaturas de uso general, locuciones adverbiales y otras palabras gramaticales que no se han recogidos al principio del proceso, confiamos en que ExtraTerm nos proporcione una selección de candidatos de términos mejorada.

2. Reducción del número máximo de unidades que forman un candidato a término, pasando de unidades de longitud de 10 palabras a 6, ya que en nuestro estudio no hemos encontrado ningún término que exceda este número.

3. Configuración de ExtraTerm con el fin de que extraiga candidatos que se encuentre en un intervalo de porcentajes más acotado, con el fin de evitar el exceso de ruido producido por el extractor.

Finalmente, con la aplicación de todas estas correcciones, confiamos en que esta herramienta pueda hacer más cómodo el trabajo cotidiano del traductor y terminólogo, ya que forma parte de un paquete de traducción asistida y ofrece posibilidades de integración con otros programas.

5. FUTUROS TRABAJOS

Como ya se ha comentado anteriormente, y a pesar de la revisión a la que se someten los documentos antes de introducirlos en el corpus, hemos detectado errores ortotipográficos al realizar la validación de los candidatos a término; errores derivados tanto de los documentos originales como de la digitalización de los mismos. Por este motivo, la primera medida a tomar es realizar una revisión exhaustiva del corpus TXTCeram con el fin de enmendar los errores ortotipográficos que hemos hallado a lo largo de nuestro análisis.
Además, somos conscientes de que se ha realizado, en este estudio, un análisis de candidatos a término de forma aleatoria. No obstante, estimamos que sería muy interesante realizar un estudio por orden porcentual, es decir, tomar como referencia la puntuación con la que aparecen los candidatos a término con el fin de determinar entre qué márgenes se sitúan la gran mayoría de los términos. Consideramos que con ello se podría evitar, en gran medida, la presencia de la gran cantidad de ruido con la que nos hemos encontrado; aunque

no por ello se impediría cierto silencio que causaría la no aparición tanto de términos con alta como con baja puntuación.

Asimismo, pensamos que podría ser muy útil realizar una investigación en la que se compararan la extracción automática con ExtraTerm y la extracción manual tanto por parte de especialistas en la materia, como por terminólogos, tal y como han hecho otros investigadores en la materia (Fulford, 2001). Lo que ayudaría, aún más, a determinar el grado de eficacia y evaluar tanto el ruido como el silencio que pueda generar la herramienta de extracción automática objeto de nuestro estudio.

6. BIBLIOGRAFÍA

• Alcina Caudet, Mª Amparo y Victoria Soler Puertes (2007): "Procedimientos y programas informáticos para la extracción automática de términos: estudio preliminar en un corpus del ámbito de la cerámica" en *X Simposio Internacional Comunicación Social, Centro de Lingüística Aplicada*, Santiago de Cuba.

• Cabré Castellví, M. Teresa, Rosa Estopa Bagot, Jordi Vivaldi Palatresi (2001): "Automatic term detection: A review of current systems" en D. Bourigault, C. Jacquemin, M-© L'Homme (eds.) *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins (en prensa).

• Estopà Bagot, Rosa (1999): "Eficiencia en la extracción automática de terminología" en Perspective: *Studies in Traductology*, Volume 7:2.

• Estopà Bagot, Rosa (2000): "Extracción de terminología: elementos para la construcción de un extractor" en *Actas de la Escuela de Terminología de Invierno de la Universidad Sao Paulo*, Brasil (en prensa).

• Fulford, Heather (2001): "Exploring terms and their linguistic environment in text: A domain-independent approach to automated term extraction" en *Terminology* 7:2, John Benjamins Publishing Company.

• Lemay, Chantal, Marie-Claude L'Homme and Patrick Drouin (2005): "Two methods for extracting "specific" single-word terms from specialized corpora" en I*nternational Journal of Corpus Linguistics* 10:2. Amsterdam: John Benjamins.

• L'Homme, Marie-Claude (2005): "Conception d'un dictionnaire fundamental de l'informatique et de l'Internet: Sélection des entrées" en *Le Langage et l'homme*, vol. XXXX, nº1, Université de Montreal.

• Soler Puertes, Victoria, Mª Amparo Alcina Caudet y Anna Estellés Palanca (2006): "La digitalización de textos para la elaboración de un corpus lingüístico electrónico: una experiencia de trabajo en equipo con estudiantes" en *X Jornades de Traducció i Interpretació a Vic: Tecnologies a l'abast*, Universitat de Vic.

• Vivaldi, Jorge y Horacio Rodríguez (2001): "Improving term extraction by commbining differents techniques" en *Terminology* 7:1, John Benjamins Publishing Company.