Volume 11, No. 4
October 2007

Antoni Oliver is a teacher in the Languages and Cultures department of the Open University of Catalonia (UOC). He has a PhD in Linguistics and his main areas of research interest are machine translation and assisted translation tools. He is also the academic coordinator of the postgraduate course on "Translation and technologies" at the UOC.

Antoni can be reached at aoliverg@uoc.edu.

Joaquim Moré works as a computational linguist at the Linguistic Service of the Open University of Catalonia (UOC). He has a master degree in English Philology and in Computational Linguistics.

Joaquim can be reached at jmore@uoc.edu.

Mercè Vázquez works as a linguist at the Linguistic Service of the Open University of Catalonia (UOC). She has a degree in Catalan Language and Literature from the University of Barcelona and also a degree in Information and Communication Sciences from the Open University of Catalonia (UOC)

Mercè can be reached at mvazquezga@uoc.edu.

Front Page

Select one of the previous 41 issues.

Select an issue:

- **Index 1997-2007**
- **TJ Interactive: Translation Journal Blog**

Translator Profiles

- On the Importance of Schmoozing
  by Alexandra Russsell-Bitting
- Standing Tall in the Profession: Interview with Alexandra Russell-Bitting
  by Verónica Albin

The Profession

- The Bottom Line
  by Fire Ant & Worker Bee

Translators Around the World

- Maltese Translation in Transition
  by Janet Mallia

TJ Cartoon

- Great Moments in Languages — Gift from Heaven
  by Ted Crump

Translation Theory

Translators' Tools

Linguoc LexTerm:

una herramienta de extracción automática de terminología gratuita

Antoni Oliver, Mercè Vázquez, Joaquim Moré

- **Introducción**

En este artículo presentamos LexTerm, una herramienta de extracción automática de terminología que es gratuita y de código abierto. Con esta herramienta se facilita la selección de los términos más relevantes que deben tener un equivalente de traducción consistente. Muchos traductores y algunas agencias de traducción realizan esta tarea todavía a mano. Aunque existen herramientas de extracción automática de terminología, éstas no son gratuitas y son programas propietarios, características que influyen en la implantación y consolidación del uso de estos programas por parte de los profesionales de la traducción. La gratuidad y el hecho de que LexTerm sea de código libre, con la ventaja de que puede ser adaptada a las necesidades específicas del usuario, pueden ser determinantes en la consolidación de un método de trabajo semi-automático, consistente en la extracción automática de candidatos a términos y la selección manual de los términos junto con sus equivalentes de traducción. LexTerm agiliza la gestión terminológica de los proyectos de traducción, y sobretodo la hace más rentable, especialmente si la comparamos con la gestión basada en la extracción manual.

LexTerm se puede descargar libremente de www.linguoc.cat.

- **La extracción automática de terminología**

La extracción de terminología es el proceso mediante el cual se seleccionan de un texto o conjunto de textos unidades candidatas a constituir términos. No hay que confundir la extracción de terminología con la creación de un glosario terminológico a partir de un texto y de una base de datos terminológica. En el caso de la extracción automática de terminología, intentamos descubrir los términos más relevantes sin conocer previamente estos términos. En cambio, en el segundo caso, buscamos qué términos de una base de datos terminológica están presentes en un determinado texto y por lo tanto los posibles términos son conocidos *a priori*.

LexTerm está diseñado para ser también utilizada por cualquier profesional que tenga que elaborar documentos con mucho contenido terminológico.

La extracción automática de terminología es una aplicación de la lingüística computacional muy interesante para la actividad del traductor tanto en la fase de preparación de un proyecto, como posteriormente, una vez finalizado. En la fase de preparación de un proyecto podemos descubrir los términos más relevantes y unificar los equivalentes de traducción. Esta posibilidad es especialmente interesante en proyectos grandes donde participan diferentes colaboradores. También es interesante extraer terminología a partir de proyectos ya finalizados, para poder recopilar entradas terminológicas que se puedan utilizar en futuros proyectos.

Existen dos técnicas principales para la extracción automática de terminología:

1. La técnica estadística que se basa principalmente en la frecuencia de aparición de una serie de combinaciones de palabras
2. La técnica lingüística que se basa principalmente en detectar patrones de categorías morfológicas

- ***Técnica estadística***

La información básica que utilizan los sistemas de extracción automática de terminología de tipo estadístico es la frecuencia de aparición de una serie de combinaciones de palabras. Los sistemas de extracción de terminología estadísticos trabajan con ***n-gramas*** de palabras. Los *n-gramas* de palabras son combinaciones de *n* palabras consecutivas. Por ejemplo, en la frase:

Nuestro sistema de gestión empresarial incluye un programa de facturación y una base de datos de recursos humanos.

Los 1-*gramas* presentes en el texto son: *Nuestro, sistema, de, gestión, empresarial, incluye, un, programa, facturación, y, una, base, datos, recursos, humanos.*

Los 2-*gramas* son: *Nuestro sistema, sistema de, de gestión, gestión empresarial, empresarial incluye, incluye un, un programa, programa de, de facturación, facturación y, y una, una base, base de, de datos, datos de, de recursos, recursos humanos.*

Los 3-*gramas* son: *Nuestro sistema de, sistema de gestión, de gestión empresarial, gestión empresarial incluye, empresarial incluye un, incluye un programa, un programa de, programa de facturación, de facturación y, facturación y una, y una base, una base de, base de datos, de datos de, datos de recursos, de recursos humanos.*

Y de esta manera sucesivamente hasta el orden *n* deseado. Los candidatos a término se encontrarán entre estas combinaciones (por ejemplo, *gestión empresarial* o *base de datos*) aunque también habrán muchos elementos sin ningún interés desde el punto de vista terminológico (por ejemplo, *programa de* o *incluye un programa*). Para poder seleccionar las combinaciones con una mayor probabilidad de constituir términos se puede hacer uso de listas de palabras vacías o *stop-words*. Las palabras vacías cuando hablamos de extracción de terminología son una serie de palabras (en su mayoría funcionales) que no pueden estar en ciertas posiciones de la entrada terminológica (normalmente las posiciones extremas, es decir, primera y última). Por ejemplo, si nuestra lista de palabras vacías para el castellano está compuesta por las palabras *nuestro, nuestra, nuestros, nuestras, de, uno, una, unos, unas, y*... y eliminamos los *bigrams* y *trigrams* que tienen en posición extrema una de estas palabras, la lista de candidatos se reduce a:

Los 2-*gramas* son: *gestión empresarial, empresarial incluye, recursos humanos.*

Los 3-*gramas* son: *sistema de gestión, gestión empresarial incluye, incluye un programa, programa de facturación, base de datos, datos de recursos.*

Si ahora trabajamos también con la frecuencia, probablemente dentro de unos textos especializados en empresa saldrá más veces *gestión empresarial* que *gestión incluye*. De esta manera se puede extraer una lista de candidatos a constituir términos, que tendrá que ser revisada manualmente.

Estos sistemas tienen dificultades para detectar los términos formados por una única palabra. Esta dificultad radica en el hecho de que el cálculo de todos los *unigrams* (*n-grams* con *n* = 1) supone todas las palabras de los textos analizados. Si filtramos por palabras vacías, obtendremos todas las palabras menos las vacías y el resultado no se parecerá en absoluto a una extracción de terminología.

- ***Técnica lingüística***

Las técnicas lingüísticas de extracción de terminología se basan en la detección de patrones morfológicos. El paso previo, por lo tanto, a la extracción de terminología es el etiquetado morfosintáctico del texto o textos. El etiquetado de textos consiste en añadir información morfológica a cada palabra del texto. Por ejemplo, si tomamos la misma frase que la utilizada para ejemplizar la técnica estadística: *Nuestro sistema de gestión empresarial incluye un programa de facturación y una base de datos de recursos humanos.* El análisis morfosintáctico daría un resultado como el siguiente:

> **Nuestro** {*nuestro* DP1MSP} **sistema** {*sistema* NCMS000} **de** {*de* SPS00} **gestión** {*gestión* NCFS000} **empresarial** {*empresarial* AQ0CS0} **incluye** {*incluir* VM1P3S0} **un** {*uno* DI0MS0} **programa** {*programa* NCMS000} **de** {*de* SPS00} **facturación** {*facturación* NCFS000} **y** {*y* CC} **una** {*uno* DI0FS0} **base** {*base* NCFS000} **de** {*de* SPS00} **datos** {*dato* NCMP000} **de** {*de* SPS00} **recursos** {*recurso* NCMP000} **humanos** {*humano* AQ0MP0} **.** {. Fp}

Para poder realizar este etiquetado es necesario disponer de un etiquetador o tagger.[1] Una vez etiquetado el texto, la extracción de terminología consistirá básicamente en una búsqueda de patrones que sean típicamente terminológicos. Algunos patrones posibles para el castellano podrían ser:

> NC AQ: substantivo - adjetivo. Detectaría: *gestión empresarial*, *recursos humanos*
>
> NC SP NC: substantivo - preposición - substantivo. Detectaría: *sistema de gestión*, *programa de facturación*, *base de datos*

Evidentemente, el hecho de cumplir uno de estos patrones no querrá decir necesariamente que se trate de una entrada terminológica. Una vez detectados los posibles candidatos, se hace una elección por frecuencia de aparición y una revisión manual. Esta metodología de extracción también presenta dificultades para la detección de términos monopalabra, ya que el patrón más habitual sería simplemente "substantivo" y detectaría todos los sustantivos del texto de entrada.

- **La búsqueda automática de equivalentes de traducción**

Si disponemos de un corpus paralelo o de una memoria de traducción es posible, además de detectar una serie de candidatos a término, encontrar automáticamente el equivalente de traducción más probable. De esta manera podemos elaborar de una manera rápida y eficiente bases de datos terminológicas bilingües.

La búsqueda automática de equivalentes funciona de la siguiente manera. Tomamos un candidato a término en la lengua A y seleccionamos un subconjunto del corpus paralelo consistente en todas llos segmentos que contienen ese término. Nos quedamos únicamente con los segmentos correspondientes a la lengua B y realizamos una extracción de terminología (habitualmente de tipo estadístico) sobre estas oraciones. El candidato a término más frecuente en estas oraciones será el equivalente de traducción más probable del término seleccionado. Esto es así ya que se espera que en todas, o la mayoría, de las oraciones en la lengua B de este subconjunto aparezca el equivalente de traducción del término seleccionado.

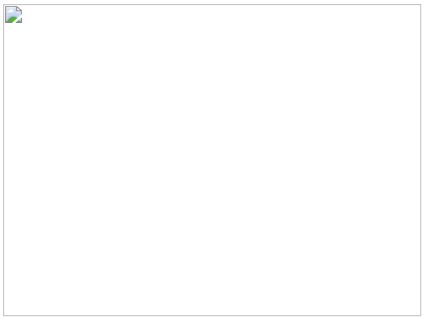
- **Linguoc LexTerm**

El programa LexTerm es un programa de extracción automática de terminología de tipo estadístico gratuito, de libre distribución y de código abierto. LexTerm pone al alcance de cualquier traductor, terminólogo, empresa o institución la posibilidad de crear glosarios terminológicos de una manera rápida y eficiente.

LexTerm está desarrollado íntegramente en Perl y por lo tanto es un programa multiplataforma. Se distribuye con el código Perl que se puede ejecutar en cualquier ordenador que tenga instalado un intérprete de Perl, y también en versión ejecutable para Windows. La versión ejecutable para Windows funcionará por cualquier ordenador que incorpore este sistema operativo tenga o no instalado el intérprete de Perl.

LexTerm permite la extracción automática de terminología y la búsqueda automática de equivalentes de traducción. En la siguiente captura de pantalla podemos observar el resultado para un corpus paralelo inglés - castellano.

Cuando se busca un equivalente de traducción siempre selecciona el más probable, pero es posible visualizar todos los posibles candidatos por orden de probabilidad. LexTerm también permite visualizar los contentos donde aparece el término original y traducido, esto puede ser de utilidad en caso de duda.



Una vez revisada la lista de candidatos y de equivalentes el resultado de la extracción se puede exportar a un formato de texto separado por tabuladores. A partir de este formato se pueden introducir los términos en un sistema de gestión de terminología o importarlos en una base de datos terminológica de alguna herramienta de traducción asistida.

- **Ejemplo de resultados**

Después de presentar el funcionamiento de LexTerm, pasamos a describir un caso práctico para mostrar el rendimiento que podemos obtener de esta herramienta. Para ello, hemos utilizado CRATER (*Corpus Resources and Terminology Extraction*), un corpus trilingüe (español, inglés y francés) de un millón de palabras, etiquetado manualmente y alineado a nivel de frase. Este corpus se ha constituido a partir de una colección importante de textos del ámbito de las telecomunicaciones, específicamente contiene el manual de la *International Telecommunications Union* (CCITT).

Con LexTerm hemos realizado la extracción de candidatos a término del corpus poniendo como límite *3-gramas* y utilizando un fichero de *stop-words* para filtrar las palabras vacías de contenido o palabras funcionales (artículos, preposiciones, etc.) que se encuentran en el corpus. A partir del resultado obtenido, hemos seleccionado el conjunto de candidatos que han aparecido en el corpus hasta la frecuencia 100. En este primer paso hemos obtenido un total de 333 candidatos a término. De este conjunto de candidatos, destacamos que entre los treinta primeros que tienen más presencia en el corpus, hemos recuperado siete que son posibles candidatos a término y son los siguientes: *earth station*, *signal unit*, *signalling link*, *signalling system*, *signalling point*. *forward signal*, *link layer*.

Este primer resultado obtenido con la herramienta de extracción automática lo hemos filtrado para poder eliminar los candidatos que pertenecen a la lengua general. Este filtraje lo hemos realizado a partir de un corpus de la lengua general que consta de un millón de palabras y ajustando el umbral de frecuencia, parámetro que permite controlar el nivel de exigencia para considerar si un candidato a término pertenece a la lengua general o al vocabulario específico. Al realizar este filtraje semiautomático, el número de candidatos ha pasado de 333 a 134, es decir, hemos reducido en un 60% los candidatos a término utilizando un umbral de frecuencia de 40. En este caso, entre los treinta primeros candidatos que tienen más presencia en el corpus, hemos recuperado ocho posibles candidatos a término y son los siguientes: *earth station*, *signal unit*, *signalling link*, *signalling system*, *signalling point*. *forward signal*, *link layer*. *backward direction*.

Al final del proceso, revisamos manualmente cada uno de los candidatos y hacemos una selección de candidatos a partir de la consulta del contexto en el que se encuentran. Este paso es más lento pero muy útil para hacer la selección definitiva de candidatos; en este caso, al realizar esta revisión manual hemos obtenido un resultado de 85 posibles candidatos, entre ellos destacamos los quince que tienen más presencia en el corpus: *earth station*, *signal unit*, *signalling link*, *signalling system*, *signalling point*, *forward signal*, *link layer*, *backward direction*, *signalling network*, *data link layer*, *satellite service*, *circuit group*, *link set*, *signalling equipment*, *transit exchange*.

- **Mejoras en la selección de candidatos**

Hasta el momento, la extracción de términos por parte de LexTerm tiene varios aspectos que son susceptibles de ser mejorados. En primer lugar, una vez se han filtrado los *n-gramas* según el criterio de aparición de *stop-words*, el filtro definitivo de posibles candidatos a término se basa en un criterio cuantitativo (la frecuencia de aparición de los *n-gramas*) y no en un criterio cualitativo que sea próximo al criterio que un humano aplicaría a la hora de realizar una selección manual. El criterio basado en la frecuencia es problemático. Por un lado, hay que seleccionar manualmente los candidatos a término de una lista larga de *n-gramas* que superan una determinada frecuencia. Muchos de ellos no se aprovechan y, sin embargo, el tiempo empleado en revisarlo es considerable. Una forma de acortar la lista de candidatos es subiendo el umbral de frecuencia de aparición, pero entonces muchos términos relevantes quedan fuera por no superar este umbral. Por esta razón hemos estado trabajando en el establecimiento de un filtro más cualitativo a partir del cual el resultado de la extracción sea una lista con candidatos más relevantes.

El filtro que estamos desarrollando se basa en la siguiente asunción: un *n-grama* que tenga una baja frecuencia en un corpus de vocabulario general pero que, por el contrario, tenga una frecuencia considerable en el corpus de especialidad sobre el cual se hace la extracción es un posible candidato a término. Para probar la validez de esta asunción estamos desarrollando un programa que filtra, de momento, palabras (*uni-gramas*) según este criterio. El programa requiere tres parámetros. El primero es una lista de palabras con sus correspondientes frecuencias, extraídas del corpus de especialidad. El segundo parámetro es el fichero que contiene el corpus de vocabulario general (o bien un fichero con las frecuencias de las palabras de este corpus). Finalmente, el tercer parámetro es un valor numérico que indica el nivel de exigencia para considerar una palabra como de vocabulario general. Este valor numérico corresponde al umbral de frecuencia que una palabra en el corpus general debería sobrepasar para ser considerada una palabra del vocabulario general y no de especialidad. El resultado del programa es una lista de palabras del corpus de especialidad que no superan el umbral que establece la frecuencia mínima que debería tener como para ser considerada una palabra del vocabulario general. El umbral no es estático, el usuario puede variarlo. Sabiendo que cuanto más pequeño sea el umbral más palabras serán descartadas, el usuario puede ir regulando el umbral hasta obtener una lista de candidatos que le parezca lo suficientemente significativa. Con el uso de estos filtros para seleccionar términos de distintas especialidades por parte de varias personas, es de esperar que podamos establecer un nivel de umbral por defecto que sea lo suficientemente productivo. Si los datos obtenidos indican la bondad de la asunción, integraremos este programa en LexTerm.

Otra mejora que debería tenerse en cuenta también afecta al filtraje. Hasta el momento, con LexTerm es posible recuperar una combinación de *n-gramas* que, a su vez, aparecen combinados entre si, es decir, para un candidato a término como "echo control device" el programa recupera las combinaciones siguientes: "echo control", "control device" y "echo control device". Cuando se lleva a cabo la revisión manual de todos los candidatos, es necesario consultar el contexto en el que se encuentran para poder discernir qué candidato es óptimo para ser escogido. En este sentido, una mejora importante sería poder disponer de todas estas combinaciones agrupadas, para poder discriminar cuál es relevante. La agrupación de los candidatos también sería útil para poder detectar y eliminar de forma rápida las variantes que encontramos de un mismo candidato: mayúsculas y minúsculas, singular y plural, etc. Así pues, esta mejora facilitaría la revisión manual de la lista de candidatos, ya que en estos momentos conlleva tener que realizar consultas minuciosas del contexto en el que éstos se sitúan.

- **Conclusiones**

Actualmente la tarea de detección de un candidato en un texto de especialidad se realiza de forma manual, lo que comporta una elevada inversión de tiempo y una alta implicación de personal especializado en el proceso de trabajo. Ante esta situación, el uso de LexTerm permite incorporar un salto cualitativo en el proceso de detección de candidatos a término de un texto de especialidad, puesto que la detección se realiza de forma automática. Con LexTerm, es posible acelerar la detección de candidatos en la fase inicial del proceso, lo que representa poder reducir el tiempo que se emplea en el marcaje manual de éstos y permite disponer al final del proceso de un mayor número de términos escogidos. Este salto cualitativo representa un significativo avance en la tarea de normalización de términos en una lengua.

De todos modos, la selección de candidatos hasta el momento es una tarea todavía larga y es inevitable que no se detecten algunos términos importantes de la especialidad porque no superan un umbral de frecuencia de aparición. Por esta razón, nuestro objetivo en el futuro es aplicar además un criterio cualitativo que filtre los candidatos a término, de modo que el número de candidatos a revisar sea menor y que la lista tenga una cobertura de términos significativos de la especialidad lo más amplia posible.

Aunque hemos presentado las utilidades de la herramienta para un traductor, LexTerm está diseñado para ser también utilizada por profesores en la confección de materiales docentes y por cualquier profesional que tenga que elaborar documentos con mucho contenido terminológico y cuya denominación en una o más lenguas debe ser absolutamente rigurosa.

- **Referencias**

CRATER (*Corpus Resources and Terminology Extraction*) (1995)
http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html

[1] Existen algunas opciones gratuitas como por ejemplo el Freeling http://garraf.epsevg.upc.es/freeling/demo.php)