Ponencia Nº: 16
Título: Multilingüismo en las instituciones europeas: herramientas y recursos para la integración del catalán

Autor(es): Borrell, Cristina; Oliver, Antoni

Tema: "La traducción es muy cara y lenta", o mentiras del monolingüismo

Correo electrónico: aoliverg@uoc.edu, cborrellc@uoc.edu

Página Web que pueda completar la temática tratada: http://multilingualismchair.uoc.edu/

Resumen en inglés

The Linguamón-UOC Chair in Multilingualism of the Universitat Oberta de Catalunya has developed an automatic extractor of terminology, which is freely distributed, multi-platform and adaptable to users' needs. One of its most important useful applications is the elaboration of glossaries, both monolingual and multilingual, based on a set of documents. This tool can be very helpful for translators and translation agencies.

The system automatically extracts the term candidates from the texts introduced by the user. Later, a specialist reviews the list of lexical units to verify the result.

In the Chair in Multilingualism we have applied this tool to the expansion the Eurovoc glossary with its Catalan version. From the entries in Spanish and a bilingual Spanish - Catalan parallel corpus we have extracted the equivalents for this language.

With this project we try to demonstrate that having the suitable resources can remarkably reduce the translation costs.

Resumen en francés

La Chaire Linguamón-UOC de Multilinguisme de l'Universitat Oberta de Catalunya a développé un extracteur de terminologie automatique de libre distribution, multiplateforme et adaptable aux besoins des utilisateurs. Son utilité la plus remarquable est l'élaboration de glossaires, monolingues ou multilingues, sur la base d'un ensemble de documents. Cet outil peut être très utile pour les traducteurs et les agences de traduction.

Le système extrait automatiquement les candidats à terme des textes introduits par l'utilisateur. Ensuite, un spécialiste révise la liste d'unités lexicales pour vérifier le résultat.

Dès la Chaire de Multilinguisme nous avons appliqué cet outil pour élargir le glossaire Eurovoc avec sa version catalane. À partir des entrées en espagnol et d'un corpus parallèle espagnol - catalan nous avons extrait les équivalents pour cette langue.

Avec ce projet nous prétendons démontrer qu'en disposant des ressources adéquates on peut réduire notamment les coûts de traduction.

1. INTRODUCCIÓN

La Cátedra de Multilingüismo Linguamón-UOC de la Universitat Oberta de Catalunya ha desarrollado un proyecto utilizando herramientas de libre distribución para crear recursos terminológicos relacionados con la Unión Europea que incluyan el catalán. Actualmente son

muy pocos los documentos oficiales que se traducen a esta lengua, aduciendo, entre otras cosas, el supuesto elevado coste que ello conllevaría. Con este proyecto pretendemos demostrar que disponiendo de los recursos adecuados y utilizando herramientas de traducción asistida y automática, los costes de traducción se reducen notablemente y permiten que, con un presupuesto asumible, se pueda respetar y aprovechar la riqueza cultural que nos brinda la diversidad lingüística.

El 2008 es el año europeo del diálogo intercultural y el año internacional de las lenguas. Actualmente, la UE está compuesta por 27 Estados miembros[1]. Es precisamente en este marco cultural incomparable que se debería sacar todo el provecho posible del patrimonio lingüístico de qué disfrutamos.

En la Unión Europea actual hay 23 lenguas oficiales[2]. A pesar de esto, pocos estados son realmente monolingües, si bien es cierto que sólo unos pocos reconocen su multilingüismo ante la Unión Europea[3]. España sólo reconoce como lengua oficial en la UE el español[4]. En cambio, Bélgica reconoce el francés, el neerlandés y el alemán.

La documentación oficial gestionada por la Unión Europea se limita a las 23 lenguas oficiales en el mejor de los casos. Los documentos de derecho primario tienen que estar redactados en todas las lenguas oficiales en la UE, para respetar la transparencia, democracia y legitimidad que prodiga la Comunidad. A pesar de esto, todavía muchos europeos se encuentran limitados a la hora de participar en las instituciones europeas, ya que gran parte de la riqueza multicultural y multilingüe que existe dentro de las fronteras de la UE queda olvidada.

Con el fin de gestionar todo el volumen de documentos multilingües y ayudar tanto al ciudadano como a los trabajadores comunitarios, la Unión Europea se ha dotado de unos servicios lingüísticos especializados[5]. Para el interés de este proyecto, pero, mencionaremos sólo la traducción en la Comisión Europea, el servicio más representativo y uno de los más grandes del mundo.

La Dirección general de Traducción de la Comisión Europea (DGT) cuenta con cerca de 2.500 trabajadores, sin tener en cuenta los trabajadores autónomos, y produce alrededor de 1,3 millones de páginas anuales. Las unidades de traducción se organizan en departamentos lingüísticos, uno para cada lengua de la UE. En los últimos años el proceso de traducción se ha ido perfeccionando, especialmente con la incorporación de la tecnología ofrecida por las memorias de traducción, de modo que los traductores pueden optimizar su esfuerzo. A grandes rasgos, cada traductor utiliza para el desarrollo de su tarea:

- Terminología (diccionarios, glosarios, bases de datos terminológicas, etc.);
- Documentos paralelos y de referencia;

[1] Alemania, Austria, Bélgica, Bulgaria, Chipre, Dinamarca, Eslovaquia, Eslovenia, España, Estonia, Finlandia, Francia, Grecia, Hungría, Irlanda, Italia, Letonia, Lituania, Luxemburgo, Malta, Países Bajos, Polonia, Portugal, Reino Unido, República Checa, Rumanía, Suecia.

[2] Alemán, búlgaro, checo, danés, eslovaco, esloveno, español, estoniano, finlandés, francés, griego, húngaro, inglés, irlandés, italiano, letón, lituano, maltés, neerlandés, polaco, portugués, rumano y sueco.

[3] En la actual Unión Europea, los Estados miembros que consideran oficiales más de una lengua són Bélgica, Finlandia, Irlanda, Luxemburgo, Malta y Chipre.

[4] Como veremos más adelante, aunque en territorio español se hablan otras lenguas, como el catalán, el gallego o el vasco, éstas sólo disfrutan de una condición restringida, y en ningún caso oficial.

[5] Toda la información referente a las lenguas de Europa y los distintos servicios lingüísticos, tanto los internos como los «abiertos», se puede consultar en el portal de la UE *Europa y las lenguas.*

- Acceso a textos previamente traducidos (especialmente mediante memorias de traducción);
- Personal de administración encargado de cumplir las tareas de pre y postedición.

Otro de los aspectos interesantes de la DGT es el conocido como "antenas de multilingüismo" que, en cierto modo, promueven la riqueza de la diversidad lingüística de Europa. La función de estas ramificaciones consiste en adaptar las comunicaciones de la Comisión a las lenguas de la región y hacer de puente entre la población local y las instituciones europeas.

Gracias a los acuerdos administrativos bilaterales entre las instituciones y el gobierno español[6], desde el 13 de junio de 2005 los catalanohablantes podrán dirigirse a las instituciones en su lengua. Lo podrán hacer en los ámbitos de las comunicaciones escritas al Consejo de la Unión Europea y a la Comisión Europea, mediante un organismo intermedio designado por el gobierno estatal que se encargará de traducir el texto (del catalán al español), y que enviará la traducción a la institución correspondiente.

2. HERRAMIENTAS

El Eurovoc es un tesauro multilingüe de los ámbitos de actividad de la Unión Europea. Actualmente se puede consultar en alemán, búlgaro, checo, croata, danés, eslovaco, esloveno, español, estoniano, finlandés, francés, griego, húngaro, inglés, italiano, letón, lituano, neerlandés, polaco, portugués, rumano y sueco. Es el único recurso comunitario que se puede descargar desde su página web.

Partiendo de esta información, podremos elaborar un glosario multilingüe. Para empezar, hemos seleccionado las lenguas "disponibles" que más nos interesan. El inglés es la lengua en la que se produce más documentación comunitaria, seguida de relativamente lejos del francés y el alemán. Lo más lógico, pues, es elegir el inglés (por la cantidad de documentos originales) y el francés, que es una lengua románica y siempre será más próxima al catalán, nuestra lengua de llegada, que el alemán.

Asimismo, también consideraremos el español. Esta es una de las lenguas más próximas al catalán, y en cualquier caso, siempre será más probable encontrar documentos paralelos en esta combinación (catalán-español). Sin ir más lejos, y quedándonos incluso dentro del ámbito institucional, la *Generalitat de Catalunya* ofrece su Diario Oficial[7] tanto en catalán como en español.

Todo este potencial lingüístico no se puede aprovechar, combinar y adaptar si no se dispone de ninguna herramienta que nos ayude a manipularlo. Para ello se ha utilizado un extractor de terminología automático que nos permite obtener los equivalentes en catalán a partir de los términos del Eurovoc en castellano. Esto es precisamente lo que hace *The Free Terminology Extraction Suite*, el conjunto de herramientas utilizado para realizar este paso del proyecto.

Gracias a esta colección de instrumentos se puede seleccionar, manipular y recuperar la terminología automáticamente, lo cual es de gran ayuda para el traductor. Además se

[6] Conclusiones del Consejo relativas al uso oficial de otras lenguas (2005/C 148/01).

[7] DOGC <http://www.gencat.net/dogc/>.

distribuye con una licencia de software libre, para que el usuario pueda bajárselo libremente[8], utilizarlo, distribuirlo o modificarlo según sus necesidades. Los instrumentos de este paquete están programados en Perl, de manera totalmente modular con tal de facilitar, si procede, su modificación y adaptación. El conjunto funciona para plataformas diferentes, tanto para Linux como para Windows.

Esta herramienta permite:

- extraer los candidatos a término directamente del texto;
- ayudar en la búsqueda de la mejor traducción para un término;
- construir glosarios terminológicos automáticamente, tanto monolingües como multilingües;
- crear bases de datos terminológicas de ámbitos específicos;
- construir automáticamente listas de unidades léxicas de un ámbito específico.

Si el *input* del conjunto de herramientas de extracción terminológica es un corpus monolingüe, el *output* resultante será una lista monolingüe de candidatos a término. Esta lista tiene que ser revisada manualmente por un lingüista, que deberá confirmar o rechazar las propuestas seleccionadas automáticamente por el extractor. Si, en cambio, el *input* está compuesto por documentación en varias lenguas, aparte de la lista se conseguirá también un recurso terminológico multilingüe, porque las herramientas permiten detectar de manera automática los posibles equivalentes de traducción de un candidato.

Este paquete de herramientas lingüísticas consiste esencialmente en un extractor de terminología y en un dispositivo de búsqueda automática de equivalentes de traducción, programados esencialmente a partir de métodos estadísticos.

1. El extractor automático de terminología funciona a partir de un documento, un conjunto de documentos o un corpus paralelo y listas de palabras vacías o funcionales para las lenguas implicadas. El método estadístico identifica candidatos a término basándose en su frecuencia en un corpus de especialidad. El proceso, a grandes rasgos, es este: se calculan estadísticamente los *n-gramas* (secuencia de n elementos, normalmente desde n = 2 hasta una *n* determinada por el usuario; es decir, todas las combinaciones de dos palabras, de tres palabras, etc.). Lógicamente, entre estas combinaciones se encontrarán algunas que no serán relevantes, por esto es imprescindible filtrar los resultados con la lista de palabras vacías. La salida de este módulo ya será una lista de candidatos a término.
2. En el segundo caso, se coge un candidato a término en una lengua A y se selecciona un subconjunto del corpus paralelo con todos los segmentos que contengan este término. Si, a continuación, se realiza una extracción terminológica estadística en estos segmentos, el resultado obtenido debería ser el candidato a término en la lengua B. En efecto, se espera que la unidad terminológica más frecuente en los segmentos en la lengua B sea pues el término que se buscaba en la lengua A .

3. PROCEDIMIENTO

De todos los términos almacenados en el tesauro tomamos la terminología en español para buscar sus equivalentes en catalán a partir de un corpus paralelo. El corpus de entrada seleccionado es el Diario Oficial de la Generalitat de Catalunya (DOGC). El DOGC se

[8] Para descargas y más información, consultar el web <www.linguoc.cat>.

publica diariamente de lunes a viernes (no festivos) en dos ediciones equivalentes, la versión catalana y la versión española. De esta manera tenemos un corpus paralelo catalán-español de ámbito institucional, político y jurídico, lo cual lo convierte en uno de los textos catalanes oficiales que comparte más características con los documentos institucionales de la UE. A la vez, al ofrecer también la versión española, optimiza el trabajo del extractor: sólo debe buscar los términos del Eurovoc en la versión española del DOGC y devolver la versión catalana del término.

Aunque el método estadístico resuelva la búsqueda satisfactoriamente en muchos casos, no siempre acierta. Por esto, el programa presenta más de un candidato y deja al usuario la elección final. El primer resultado que ofrece es siempre el que ha aparecido más frecuentemente.

En este punto es donde empieza realmente la tarea del lingüista. Ahora se tendrán que revisar manualmente todas las propuestas hechas por el extractor, corregirlas si no son correctas o incluso proponer las traducciones reales si el extractor ha sido incapaz de ofrecer alguna. Esto sucederá en los casos en los que el término español que se pretende extraer para el catalán no aparezca en el DOGC. En cualquier caso no se puede ignorar el ahorro de tiempo y de investigación que se ha obtenido con el uso del extractor.

En caso de que el extractor no haya actuado satisfactoriamente, el experto que revise el resultado tendrá que lidiar con los tipos de problemas que mencionamos a continuación[9]:

1. A menudo, la primera propuesta para el catalán no es completa. Esto sucede sobre todo en casos de unidades terminológicas formadas por más de una palabra, que de algún modo desorientan a los sistemas estadísticos.
2. A veces, la versión española y la catalana no son equivalentes. No es frecuente, pero puede pasar que el extractor se equivoque y se confunda en la selección de términos que no pertenecen al mismo ámbito de especialidad.
3. A veces, el extractor nos propone n-gramas pertenecientes al mismo segmento que el término clave, pero que sin embargo no se corresponden con la posible traducción.
4. Es probable que las versiones en otras lenguas diferentes al español no se correspondan a la del catalán. Cabe recordar que el extractor toma el término en español para hacer su selección del catalán, por lo tanto en ningún momento tiene en cuenta las otras lenguas[10].
5. Puede darse el caso de que la versión en catalán sea el desarrollo de una sigla o viceversa. Hará falta que el experto lo tenga en cuenta y armonice criterios.

4. RESULTADOS

Para realizar los experimentos hemos utilizado un corpus paralelo castellano-catalán del DOGC que comprende todos los números desde el 3.544 (del 2/01/2002) al 5.118 (del 24/04/2008). El corpus ha sido segmentado a nivel de oración y alineado automáticamente. Consta de un total de 9.492.333 segmentos y 121.145.821 palabras en las dos lenguas.

[9] Estos son los errores encontrados tras la aplicación de primera versión del extractor, la mayoría de los cuales desaparecerán gracias a las mejoras de las versiones posteriores. Este análisis nos sirve para descubrir qué cambios son relevantes para mejorar el rendimiento del extractor.

[10] En última instancia, ello dependería de la calidad de los equivalentes del Eurovoc, ámbito que queda fuera del alcance de este estudio.

Los resultados que presentamos en primer lugar se han obtenido para un valor de incremento de *n* de ±1 y el algoritmo de búsqueda de equivalentes de traducción no se ha optimizado. De este modo el primer candidato que aparece es el que presenta mayor frecuencia de aparición en el subcorpus paralelo formado por todos los segmentos que contienen el término original. En la tabla presentamos el porcentaje de equivalentes que aparecen en primera posición, en las tres primeras y en las diez primeras posiciones.

Posición	**Porcentaje**
1	43.4
1 - 3	73.4
1 - 10	86.2

Tabla I. Resultados obtenidos con el algoritmo sin optimizar.

Analizando los resultados hemos llegado a las siguientes conclusiones:

- Cuando se obtienen diversos candidatos en las primeras posiciones con la misma frecuencia, el candidato que se presenta en primer lugar puede ser cualquiera de ellos. Esto es debido a que se utilizan *hashes* para almacenar los candidatos, y estas estructuras de datos no tienen un orden determinado.
- A menudo el primer candidato presenta una frecuencia superior al propio término original. Esto hace que en primer lugar suelan aparecer términos monopalabra incluso si el término original era multipalabra.

Para minimizar las consecuencias del primer fenómeno hemos programado una optimización del algoritmo de búsqueda de equivalentes de traducción que consiste en agrupar todas las propuestas que tienen la frecuencia más alta y aplicarles el algoritmo relativo a la distancia de edición (cuanto más parecido es el candidato al original, más posibilidades de ser el correcto tiene), los resultados mejoran así:

Posición	**Porcentaje**
1	62.2
1 - 3	82.1
1 - 10	88.2

Tabla II. Resultados obtenidos agrupando los candidatos con mayor frecuencia y aplicando la distancia de edición.

Para evitar el problema producido por candidatos con una frecuencia superior espuria hemos aplicado una optimización similar a la anterior, pero que agrupa todos los candidatos con las tres mayores frecuencias. De este modo obtenemos los siguientes resultados:

Posición	**Porcentaje**
1	74.7

1 - 3	80.4
1 - 10	88.2

Tabla III. Resultados obtenidos agrupando los candidatos con las tres frecuencias mayores y aplicando la distancia de edición.

Como se puede observar, los resultados han mejorado notablemente, especialmente en las primeras posiciones.

6. CONCLUSIONES Y TRABAJO FUTURO

En el desarrollo del proyecto se ha obtenido un glosario multilingüe de 2.531 términos utilizados en los documentos oficiales de la Unión Europea disponibles en cuatro lenguas, entre las cuales, el catalán. Por el momento, de todos los recursos lingüísticos comunitarios sólo ha sido posible procesar el Eurovoc, que es la única herramienta que ofrecía la posibilidad directa de ser descargada.

Hemos demostrado que con unos recursos básicos y herramientas sencillas se pueden generar recursos multilingües que pueden agilizar el proceso de traducción de documentación. Con las herramientas adecuadas, el proceso de generación de estos recursos es muy ágil y, por lo tanto, económico. Además, todas las herramientas utilizadas son de libre distribución y por lo tanto disponibles para toda la comunidad[11].

La metodología utilizada se puede aplicar a un gran número de lenguas. Los únicos requisitos son disponer de un corpus paralelo de la temática tratada y una pequeña lista de palabras vacías de los pares de lenguas implicadas. Se debe tener en cuenta que la metodología estadística de extracción de terminología no funciona bien para lenguas aglutinantes.

El trabajo futuro se divide en dos líneas:

- Por un lado, aplicar la metodología descrita a otros recursos terminológicos comunitarios, como por ejemplo el IATE.
- Por otro lado, continuar trabajando en la mejora y optimización de las herramientas de extracción.

[11] Las herramientas se pueden descargar desde aquí: <http://multilingualismchair.uoc.edu>.

Referencias bibliográficas

1. COMISIÓN EUROPEA. *Informe final del Grupo de Alto Nivel sobre Multilingüismo*. [en línea]. Luxemburgo: Comisión Europea, 2007. ISBN 978-92-79-06902-4. Disponible en Internet: <http://ec.europa.eu/education/languages/archive/doc/multishort_es.pdf>.

2. COMISIÓN EUROPEA. *Multilinguisme et traduction* [en línea]. Bruselas: Dirección general de Traducción de la Comisión Europea, 2002. ISBN 978-92-79-00838-2. Disponible en Internet: <http://ec.europa.eu/dgs/translation/bookshelf/brochure_fr.pdf>.

3. COMISIÓN EUROPEA. *Translating for a multilingual community* [en línea]. Bruselas: Dirección general de Traducción de la Comisión Europea, 2002. ISBN 92-79-03586-X. Disponible en Internet: <http://ec.europa.eu/dgs/translation/bookshelf/brochure_en.pdf>.

4. COMISIÓN EUROPEA. *Un reto provechoso. Cómo la multiplicidad de lenguas podría contribuír a la consolidación de Europa* [en línea]. Bruselas: Comisión Europea, 2008. Disponible en Internet: <http://ec.europa.eu/education/languages/archive/doc/maalouf/report_es.pdf>.

5. DOUE. *Acuerdo administrativo entre la Comisión Europea y el Reino de España* (2006/C 73/06) [en línea]. Bruselas: Unión Europea, 2006. Disponible en Internet: <http://eur-lex.europa.eu/LexUriServ.do?=uri=OJ:C:2006:073:0014:0015:EN:PDF>.

6. DOUE. *Acuerdo administrativo entre el Reino de España y el Consejo de la Unión Europea* (2006/C 40/02) [en línea]. Bruselas: Unión Europea, 2006. Disponible en Internet: <http://eur-lex.europa.eu/LexUriServ/site/es/oj/2006/c_040/c_04020060217es00020003.pdf>.

7. FORCADA, Mikel. *Traducción automática de código abierto: una oportunidad para las lenguas menores* [en línea] en: Ciclo de conferencias de la Cátedra de Multilingüismo Linguamón-UOC. Barcelona: Cátedra de Multilingüismo Linguamón-UOC, 2008. Disponible en Internet: <http://www.slideshare.net/mlforcada/traduccin-automtica-de-cdigo-abierto-una-oportunidad-para-lenguas-menores/>.

8. OLIVER, Antoni, VÀZQUEZ, Mercè, MORÉ, Joaquim. *Linguoc LexTerm: una herramienta de extracción automática de terminología gratuita* [en línea]. Poughkeepsie: Translation Journal, volumen 11, nº 4, 2007. ISSN 1536-7207. Disponible en Internet: <http://accurapid.com/journal/42linguoc.htm>.

9. VÀZQUEZ, Mercè, OLIVER, Antoni. *A Free Terminology Extraction Suite,* en: Translating and the computer. Londres: 29 ASLIB Information Management, 2007.

Recursos electrónicos:

10. *Translation and drafting aids in the European Union languages* [en línea]. Bruselas: Comisión Europea, 2004. Disponible en Internet: <http://ec.europa.eu/translation/index_en.htm>.

11. *Diari Oficial de la Generalitat de Catalunya (DOGC)* [en línea]. Barcelona: Generalitat de Catalunya, 2008. Disponible en Internet: <http://www.gencat.net/dogc/>.

12. *Eurovoc. Tesauro plurilingüe* [en línea]. Bruselas: Unión Europea, 2005. Disponible en Internet: < http://europa.eu/eurovoc/>.