ANEXO III. MEMORIA FINAL DE PROYECTO

UCOTERM[1]: SITIO WEB PARA LA DIFUSIÓN DE RECURSOS PARA LA TRADUCCIÓN CIENTÍFICO-TÉCNICA BASADO EN LA CLASIFICACIÓN MULTI-CLASE

UCOTERM: WEBSITE FOR THE DISSEMINATION OF RESOURCES FOR SCIENTIFIC AND TECHNICAL TRANSLATION BASED ON MULTICLASS CLASSIFICATION

Mª Azahara Veroz González, Manuel Marcos Aldón, Fuensanta Mª Guerrero Carmona,
Soledad Díaz Alarcón, Mª Cristina Toledo Báez, María del Mar Rivas Carmona,
Beatriz Marínez Ojeda, María Aurora Toscano Crespo, Jesús Claudio Pérez Gálvez,
Cristóbal Laguna Cañero, Juan Antonio Muñoz Cecilia,
José Antonio Domínguez Barragán, Pedro Ignacio Valderrábano Madrid
averoz@uco.es
Universidad de Córdoba

Received: dd/mm/yyyy Accepted: dd/mm/yyyy

Abstract

UCOTerm is the Language Engineering laboratory of the University of Córdoba (Spain) born from «UCOTerm: Website for the dissemination of scientific and technical resources for translation», an innovative teaching project whose main purpose is the creation of terminological and multilingual resources and tools for translators and interpreters. These tools are created by our students and coordinated by our teaching staff in Translation. Our main purpose is to make our students capable of applying their acquired theoretical-practical knowledge by creating instrumental processes and terminological and textual consultation tools, such as glossaries, thesauri and specialized corpora.

Keywords: Language Engineering Laboratory, specialized translation, glossaries, thesauri, corpus

Resumen

UCOTerm es el laboratorio de ingeniería documental de la Universidad de la Universidad de Córdoba que nace a partir de un primer proyecto de innovación docente «UCOTerm: Sitio web para la difusión de recursos para la traducción científico-técnica» cuya finalidad principal es la creación de recursos y herramientas documentales multilingües por parte del alumnado y dirigidas por el profesorado de Traducción e Interpretación (Grupo docente 169) en las que se aplican los conocimientos teórico-prácticos adquiridos, materializándose en la creación de procesos instrumentales de consulta terminológica y textual, tales como glosarios, tesauros y herramientas de consulta de corpora en el ámbito científico-técnico.

UcoTerm pretende ir más allá y cubrir los diferentes ámbitos especializados de la traducción, tales como el jurídico, económico, administrativo, etc.

Palabras clave: Ingeniería documental, traducción especializada, glosarios, tesauros, corpus

1. INTRODUCCIÓN

Desde hace cinco años, en las asignaturas de Documentación, Herramientas para la Traducción Profesional y Trabajo de Fin de Grado se ha planteado, como parte del proyecto educativo, la creación de recursos y herramientas documentales multilingües por parte del alumnado y dirigidas por el profesorado de estas asignaturas en las que se aplican los conocimientos teórico-prácticos adquiridos, materializándose en la creación de procesos instrumentales de consulta terminológica y textual, tales como glosarios, tesauros y herramientas de consulta de corpora en el ámbito científico-técnico.

En esta línea se hace imprescindible añadir un paso más al proyecto educativo de dichas asignaturas con el fin de, por un lado, difundir los trabajos que nuestro alumnado realiza con tanto esmero y, por otro, aumentar la motivación del mismo pues, de este modo, el alumnado obtendría una aplicación real a su trabajo y además realizaría una prospectiva social en la que la Universidad de Córdoba produciría una visibilidad profesional. Con este proyecto nos proponemos darle transferencia de carácter profesional a las investigaciones dirigidas por el profesorado y realizadas por los alumnos a través de la puesta en marcha de un web site con las herramientas, recursos de mejor calidad en el ámbito científico-técnico desarrollados por el alumnado facilitando, a su vez, la difusión del conocimiento entre profesionales e investigadores e interconectándolo con la iniciativa internacional de la Unesco (ya se ha obtenido el estándar de interoperabilidad lingüística para los vocabularios controlados de esta institución). Hemos de tener en cuenta que la elaboración de una unidad de información con recursos de este tipo, consituye un valor añadido, no solo a nuestras asignaturas y alumnos, sino a toda la comunidad científica y profesional en general pues pone a disposición de la sociedad dichas herramientas multilingües con carácter abierto facilitando la labor profesional de muchos traductores e investigadores del ámbito científico-técnico además de contribuir a la comunicación multilingüe entre ellos. De esta forma, se integra a los estudiantes en el paso obligado de formarlos como investigadores desde los comienzos de sus enseñanzas, atrayéndolos hacia los estudios de máster que se imparten en nuestra universidad.

Se pretende, pues, diseñar, una unidad de información enlazadas con las asignaturas del departamento en la que cualquier estudiante, profesor, investigador o profesional pueda acceder desde cualquier dispositivo, fijo o portátil (teléfono, tablet, portátil) con el material de mejor calidad creado por el alumnado de las asignaturas de Documentación, Herramientas Profesionales para la Traducción y Trabajo de Fin de Grado, respetando la autoría de los mismos, pero siempre con la guía y control del profesorado especializado. Y que además favorezca el desarrollo de trabajos posteriores que impliquen un cierto grado de desarrollo en el proceso de investigación, por ello la diversidad de líneas de investigación y de información englobadas en el ámbito de la Traducción impulsa a la realización de un sistema que facilite a los estudiantes y profesores la búsqueda de información sobre estas disciplinas. Los vocabularios controlados proporcionan una forma de organizar el conocimiento para facilitar su posterior recuperación. Se utilizan en los índices temáticos, tesauros, sistemas de clasificación,

[1] Website: http://www.uco.es/ucoterm/

ontologías, etc. Un vocabulario controlado exige el uso de términos predefinidos y autorizados por los diseñadores del mismo. Cada término consiste en una o más palabras que se utilizan para representar un concepto, y se seleccionan a partir del lenguaje natural. En la actualidad no existe un vocabulario controlado exhaustivo que abarque los términos relacionados con todos los campos de investigación sobre temas de traducción especializada.

La elaboración de este material favorece igualmente el aprendizaje por competencias, en particular determinadas competencias básicas y específicas que están establecidas en el plan de estudios de Traducción e Interpretación, y que detallamos a continuación: Competencia Básica 3 (CB3). Capacidad para localizar, obtener, gestionar y transmitir información, ideas, problemas y soluciones a un público tanto especializado como no especializado. Competencia Básica 5 (CB5). Desarrollo de la creatividad y capacidad de autoaprendizaje para emprender estudios posteriores con un alto grado de autonomía. Competencia Básica 6 (CB6). Capacidad para el trabajo en equipo y la toma de decisiones en contextos internacionales e interdisciplinares. Competencia universidad 2 (CU2). Conocer y perfeccionar el nivel de usuario en el ámbito de las TIC. Competencia específica 3 (CE3). Capacidad para la búsqueda y análisis de información documental y textual y aprovechamiento de la información contenida en bases de datos, diccionarios, otros soportes informáticos e Internet en el campo de la traducción. Competencia específica 7 (CE7). Ser capaz de utilizar el metalenguaje especializado y profesional. Competencia específica 16 (CE16). Ser capaz de crear y gestionar bases de datos terminológicas. Competencia específica 20 (CE20). Ser capaz de interrelacionar los distintos aspectos de la traducción e interpretación y de relacionar el conocimiento traductológico con otras áreas y disciplinas.

Consideramos pues que sitio web de acceso abierto es un recurso formativo esencial consituyendo, asimismo, un complemento didáctico para el perfeccionamiento de las destrezas de traducción e informáticas en áreas especializadas, así como la asimilación de una metodología teórico-práctica adquirida en dichas asignaturas.

2. Objetivos

Los principales objetivos fijados en el proyecto UCOTerm han sido los siguientes:

1. Difundir las herramientas y recursos de mejor calidad desarrolladas por el alumnado de las asignaturas de Documentación, Herramientas Profesionales para la Traducción y Trabajo de Fin de Grado, respetando su autoría.
2. Motivar al alumnado con la publicación de los mejores trabajos y su entronque y prospectiva en la sociedad empresarial. Iniciar al alumnado en actividades investigadoras y de transferencia del conocimiento.
3. Facilitar el aprendizaje de los conocimientosteórico-prácticos de las asignaturas de Documentación, Herramientas Profesionales para la Traducción y Trabajo de Fin de Grado, a través de la publicación de resultados. Aprender de manera estratégica, autónoma y continua.
4. Resolver problemas documentales de textos especializados.
5. Apoyar la enseñanza virtual y el empleo de herramientas virtuales.
6. Fomentar el autoaprendizaje como medio de desarrollo, innovación y responsabilidad profesional a través de la formación práctica, continua y especializada.
7. Disponer de destrezas documentales y desarrollar competencias profesionales en el uso de manuales y fuentes documentales generales y especializados.
8. Gestionar recursos de documentación para la resolución de problemas de traducción (competencia instrumental).
9. Capacitar al estudiante para el acceso, búsqueda y selección de la información, utilizando las nuevas tecnologías de la información y comunicación.
10. Movilizar conocimientos enciclopédicos, culturales y temáticos para resolver problemas documentales (competencia extralingüística).
11. Diferenciar las lenguas controlando las interferencias (competencia lingüística contrastiva). Aprender a analizar, sintetizar, razonar críticamente y tomar decisiones.
12. Integrar las TIC como herramientas que contribuyan a que el estudiante abandone el papel de sujeto receptor y pase a ser un elemento activo de su aprendizaje.
13. Facilitar la comunicación multilingüe entre profesionales e investigadores del ámbito científico-técnico.

3. Material y Métodos

La planificación, diseño y posterior implementación del soporte conceptual de este proyecto se asienta en la aplicación de los procedimientos que exponemos a continuación.

a) Fase preliminar:

Presentación, puesta en común y valoración crítica de la propuesta para este proyecto: se expusieron las diferentes ideas para la realización del proyecto. En primer lugar, se definió la temática sobre la que iba a versar este primer proyecto, si bien es verdad, que desde un inicio el equipo tenía claro que se trataba de un proyecto inicial al cual se le debía dar continuidad una vez finalizado el proyecto, ya sea en forma de otro proyecto como sitio web colaborativo. El equipo de profesores se encargaría principalmente de definir los criterios temáticos y metodlógicos para el desarrollo de herramientas para traductores, mientras que el equipo técnico se encargaría de la puesta en marcha del sitio web, de la base de datos interna y del desarrollo del buscador.

Resolución de los contenidos sobre los que se asienta y planificación de su estructura: en este primer proyecto, la temática sería científico-técnica, si bien, como hemos comentado anteriormente, se trata de un proyecto en contínuo desarrollo que pretende abarcar las principales áreas temáticas de la traducción especializada.

Selección y decisión de los soportes y recursos necesarios para su puesta en marcha de las que se ha encargado principalmente el equipo técnico.

Una vez realizada la fase premilimar del proyecto que trata de asentar las bases para el desarrollo del mismo, nos disponemos a explicar cada una de las fases que ha tenido el proyecto, si bien, muchas de ellas se han desarrollado de manera simultánea.

b) Fase I:

Ordenación y distribución de tareas a los diferentes equipos de trabajo, en adelante Subproyectos:

Subproyecto 1: Organización y coordinación de las fases del proceso y de los respectivos subproyectos.

Subproyecto 2: Selección, desarrollo y evaluación de los mejores trabajos para su publicación.

Subproyecto 3: Selección y desarrollo de las herramientas necesarias para el volcado de datos en local.

Subproyecto 4: Volcado de datos en las herramientas en local.

Subproyecto 5: Creación del sitio web.

Subproyecto 6: Publicación de los recursos en el sitio web.

Subproyecto 7: Difusión del la unidad de información web creada a través de redes de conocimiento (RedIris) y redes socioprofesionales y redes sociales (Twitter, Facebook)y ponerlo a disposición de la estructura empresarial y de las Cámaras de Comercio andaluzas.

c) Fase II: Subproyectos 2 y 3

Subproyecto 2:

Hay que tener en cuenta que el lenguaje está siempre en un proceso de cambio continuo y desarrollo, en el nivel semántico, dónde las palabras existentes adquieren significados nuevos, y en el nivel léxico, dónde aparecen nuevos conceptos y otros desaparecen o se usan con menos frecuencia. En un campo científico-técnico como el de la Traducción, pueden surgir nuevos términos como resultado de investigaciones o descubrimientos que determinan importantes variaciones en el dominio en los que se utilizan en cada una de las disciplinas implicadas en el prooceso de traducción. Estos conceptos nuevos deben ser tenidos en cuenta en la construcción de un vocabulario controlado para el dominio objeto de estudio. Para identificar los términos específicos de un dominio se han utilizado herramientas de extracción automática de términos existente, como el Stanford Named Entity Recognizer (NER), que logra muy buenos resultados en la extracción de personas, organizaciones, lugares y otros tipos de entidades generales (Finkel, Grenager y Manning, 2005). Sin embargo, la extracción de términos para la Traducción especializada requiere un modelo específico distinto, por lo que es fundamental entrenar el extractor sobre multitud de datos etiquetados que incluimos para su localización y aplicación correctas. Debido a que no disponemos de un corpus etiquetado sobre todo este este dominio, no es posible utilizar este enfoque, por lo que se han analizado todos los documentos buscando los términos que deben formar parte de nuestro vocabulario controlado y de entre estos se han definido las herramientas documentales a publicar: glosarios y tesauros[2].

a) Glosarios:

Para los glosarios se ha realizado una plantilla en excel basada en la ficha terminológica estandarizada que propone el Centre de Terminologie de Bruxelles y que a continuación se puede consultar.

	ES-EN-FR-AR
	CAMPO DE UTILIZACIÓN:
	SUBCAMPO DE UTILIZACIÓN:
	ÁMBITO DE APLICACIÓN:
ENTRADA:	
DEFINICIÓN:	
CONTEXTO 1:	
FUENTE:	
CONTEXTO 2:	
FUENTE:	
VARIANTE:	
NOTA:	
TÉRMINO EQUIVALENTE EN INGLÉS:	
TÉRMINO EQUIVALENTE EN FRANCÉS:	
TÉRMINO EQUIVALENTE EN ÁRABE	

1. Ficha en blanco del *Centre de Terminologie de Bruxelles*

[2] Aunque en un principio se propuso introducir también corpora, de momento, esta parte del proyecto se ha parado, debido a que se está estudiando el desarrollo informático más apropiado para su puesta en marcha.

Esta plantilla de excel (Figuras 2, 3 y 4) se modificó para que los glosarios pudieran ser multilingües, en principio, hasta tres lenguas (inglés, francés y español) y bidireccionales, por lo que no solo incluía los equivalentes sino también contextos y definiciones en otras lenguas. De este modo, según las características del trabajo, el alumno podía elegir entre realizar fichas monolingües con equivalentes, o fichas bilingües o multilingües.

2. Fragmento de la plantilla de excel en inglés

3. Fragmento de la plantilla de excel en francés

4. Fragmento de la plantilla de excel en español

b) Tesauros:

Para la creación de tesauros únicamente se dio la consigna de que debían estar ordenados de manera jerárquica, con sus equivalententes al lado en otro color y siguiendo las indicaciones de la *UNE-ISO 25964-1:2014: Información y documentación. Tesauros e interoperabilidad con otros vocabularios. Parte 1: Tesauros para la recuperación de la información*.

Una vez creadas las plantillas y los criterios de elaboración de los glosarios y los tesauros, se dio difusión al proyecto entre los alumnos del Grado de Traducción e Interpretación, del Doble Grado de Traducción e Interpretación y Estudios Ingleses y del Doble Grado de Traducción e Interpretación e Hispánicas. Para ello, se creó un espacio en el ENOA de Moodle para la entrega de trabajos titulado Proyecto de vocabularios controlados y tesauros (cf. Figura 5). Asimismo, este proyecto se dio a conocer a nuestros alumnos tutelados en el Trabajo de Fin de Grado, de Máster y a alumnos de otras asignaturas que en un principio no contemplamos en el proyecto inicial, pero que consideramos que podrían ser de gran utilidad, como la asignatura de Traducción de Textos Científico-Técnicos de inglés.

5. Vista del ENOA

Con este fin se inició un proceso de adquisición de fuentes basadas en los artículos y manuales para cada uno de los campos disciplinares de los que cada glosario o tesauro versaría con ello podíamos obtener una referncia adecuada y representativa de términos ya que el objetivo principal de la minería de textos es identificar relaciones y modelos en la información que permitan obtener una visión selectiva del contenido de los documentos y detectar patrones no triviales sobre el conocimiento almacenado (Berry, 2007). Con este objetivo se realizó un análisis de clusters o agrupamiento de datos que constituye la base del aprendizaje y del conocimiento. Entre otras tareas propias de este campo hemos de destacar que lo hemos empleado de forma eficiente para determinar la similitud de documentos, para optimizar el rendimiento de los sistemas de búsqueda y de recuperación de información, y para comprobar la homogeneidad de un corpus textual o en la categorización de documentos. Para ello analizamos diversos métodos de agrupamiento para comprobar cual se ajusta mejor al campo de investigación de la Traducción Especializada.

Nos impusimos como norma el objetivo de los sistemas de clasificación automática de documentos, que no es otro que la asignación de la categoría más apropiada a cada uno de los documentos de entrada, partiendo de un esquema predefinido de clasificación. Para grandes corpus textuales se suele emplear la distancia de Kullback-Leibler (KLD, Kullback-Leibler Distance) entre la distribución de probabilidad del documento que se quiere clasificar y la distribución de probabilidad de cada categoría. Diversos estudios demuestran que los resultados obtenidos con el método KLD son superiores a los logrados con tf-idf. La clasificación automática de documentos puede ser de tres tipos: supervisada, en la que existe algún mecanismo (manual o no) que proporciona información sobre la clasificación correcta de los documentos; no supervisada o

agrupamiento, realizada sin ningún tipo de referencia ni información externa; y semi-supervisada, para la que es necesario disponer de algunos documentos etiquetados correctamente.

Las técnicas de clasificación se aplican a muchos campos de estudio, como el filtrado de mensajes, identificación del idioma de un texto, análisis de sentimientos, etc. En nuestro caso se utilizó para la clasificación en categorías de información relativa al campo de la traducción especializada. Este es un ámbito que engloba áreas y disciplinas de investigación y docencia diferentes. En la actualidad no se dispone de un vocabulario de dominio específico de la Traducción Especializada que permita representar, normalizar y compartir el conocimiento en este campo. Se construyó, como paso previo a la categorización de documentos, un vocabulario controlado y dinámico de términos relacionados semántica y genéricamente que abarque este dominio de conocimiento. La clasificación textual en esta área ayudará a los investigadores en su trabajo de búsqueda, gestión y análisis documental, lo que repercutirá en la calidad y en los resultados de sus estudios. De esta forma se podrán mejorar las prácticas de investigación de forma que permitan garantizar mejores resultados y productos de la investigación, asegurando la trazabilidad de los procesos y actividades de su investigación. Posterioremente y de forma resumida hemos realizado una parametrización del aprendizaje automático del sistema de información para que los elementos introducidos puedan realicionarse por procesos estocásticos en base a sus atributos de clase y propiedades de clase en cada término participante de los lenguajes controlados.

Este aprendizaje automático lo hemos dividido como se hace de forma tradicional en supervisado y no supervisado. En el predictivo o supervisado el objetivo es crear una función capaz de predecir el valor correspondiente a los objetos de entrada después de haber visto una serie de ejemplos de entrenamiento. Se usan por tanto pares entrada-salida.

Dado un conjunto etiquetado de pares entrada-salida (X, y) se tiene que:

$$D = \{x_i, y_i\}_{i=1}^{N}$$

Siendo D el corpus de entrenamiento, y N el número de ejemplos de entrenamiento.

En el descriptivo o no supervisado solo hay entradas, y su objetivo es encontrar patrones interesantes en los datos, sin saber a priori qué buscamos.

$$D = \{x_i\}_{i=1}^{N}$$

Tenemos por tanto unos datos de entrenamiento que constan de un conjunto de vectores de entrada x pero no tenemos sus correspondientes valores de salida, sino que trata de descubrir características similares dentro de los datos.

Como se indica en (Bishop, 2006) el objetivo de los problemas de aprendizaje no supervisado puede ser el descubrimiento de grupos de ejemplos similares en los datos, conocido como *clustering* o agrupamiento, la determinación de la distribución de los datos dentro del espacio de entrada, llamado estimación de densidad, o la reducción de la dimensión de los datos pasando de un espacio de alta dimensionalidad hasta un espacio de dos o tres dimensiones, con el objetivo de poder visualizar dichos datos.

A diferencia del aprendizaje supervisado, que calcula la densidad condicional de los datos, el no supervisado calcula la densidad no condicional (Murphy, 2012). Además, x_i es un vector de características, por lo que necesitamos crear modelos de probabilidad multivariante, mientras que en el aprendizaje supervisado, y_i suele ser una variable simple que estamos tratando de predecir. Esto significa que para la mayoría de problemas de aprendizaje supervisado, podemos usar modelos de probabilidad univariante, con parámetros que dependen solo de la entrada, simplificando considerablemente el problema.

El aprendizaje no supervisado se puede aplicar en muchas más ocasiones que el supervisado, ya que no requiere que los datos estén etiquetados por un experto. Los datos etiquetados son difíciles de conseguir, y la información que suelen contener es escasa, por lo que no siempre son fiables para estimar los parámetros en modelos complejos.

Los algoritmos del aprendizaje supervisado y no supervisado se pueden combinar para poder clasificar las entradas de forma adecuada. En este aprendizaje semi-supervisado se tienen en cuenta tanto los datos etiquetados como los no etiquetados. Además, existe un tercer tipo de aprendizaje automático, conocido como aprendizaje por refuerzo (Sutton y Barto, 1998), que aprende a base de ensayo-error. Su información de entrada es el *feedback* o retroalimentación, buscando aquellas acciones con las que se obtiene mayor recompensa. En lugar de disponer de ejemplos de salidas correctas, intenta descubrirlas probando los resultados.

Conseguido esto se procedió a eleaborar un agrupamiento de documentos textuales que como técnica nos permitía elaborar correspondencias entre los diferentes elementos semánticos participantes, al ser una técnica de aprendizaje no supervisado, resultaba difícil evaluar la calidad de la salida de un método. El problema de la búsqueda de patrones en los datos ha preocupado a lo largo de la historia. Se puede encontrar gran cantidad de bibliografía referida al tema del agrupamiento.

El agrupamiento de documentos consiste en la división de los datos en grupos de objetos similares entre si, y diferentes a los objetos de los otros grupos. Cuando se trata de agrupar documentos textuales, uno de los principales problemas es el de la dimensionalidad, tanto en relación con el número de documentos que se deben analizar como con los atributos que describen cada documento. El número de características (palabras o términos) que describen los documentos de una colección de gran tamaño es tan elevado que el análisis de los mismos requiere recursos computacionales considerables.

El agrupamiento de textos se puede llevar a cabo en línea, limitado por problemas de eficiencia, o fuera de línea, descargando previamente los textos que se han de analizar. Implica la utilización y extracción de términos. Estos términos son conjuntos de palabras que describen el contenido dentro de un grupo. El método más utilizado para llevar a cabo el

agrupamiento es representar cada texto con un vector compuesto de un conjunto de términos descriptores y, a partir de ellos, se utiliza alguna función de similitud para generar grupos de textos similares (Manning y Schütze, 1999). El modelo espacio vectorial (VSM, Vector Space Model) representa los documentos textuales mediante vectores (Salton, Wong y Yang, 1975), asignando un peso a cada uno de los términos sin tener en cuenta la probabilidad de aparición del término en los documentos. Existen otros modelos de representación textual que resuelven algunos de los problemas planteados por el VSM. Posteriormente pudimos desarrollar en nuestra unidad virtual un doble buscador que implementamos para relacionar las clases y subclases semánticas entre los términos internos y con los externos, para ello empleamos como herramienta BS, Beautiful Soup es una biblioteca de Python para la extracción de datos procedentes de archivos HTML y XML. Analiza los archivos para proporcionar formas idiomáticas de navegación y búsqueda, y permite la modificación del árbol de análisis sintáctico.

La última versión es la 4.3.2, de octubre de 2013, bajo licencia MIT (Massachusetts Institute of Technology). Una vez instalado el paquete, para analizar un documento solo hay que pasarlo al constructor, a través de una cadena o como un gestor de archivos:

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(open("index.html"))
soup = BeautifulSoup("<html>data</html>")
```

BeautifulSoup, combinado con el módulo para la consulta en Google Scholar detallado en líneas posteriores, permite analizar los resultados devueltos por el buscador. Esto facilita la obtención de los documentos en pdf publicados por los investigadores sobre temas necesarios para la traducción especializada. Para ello modificamos el módulo de búsqueda en Google Scholar ya que es un buscador de Google centrado en el mundo académico, especializado en artículos científico-académicos. Este sitio web mantiene un índice de editoriales, bibliotecas, repositorios y bases de datos bibliográficas. Los resultados obtenidos al realizar una consulta incluyen datos relativos al número de citas, el enlace a libros, artículos de revistas, comunicaciones y ponencias, informes científico-técnicos, tesis, tesinas y archivos depositados en repositorios.

Lanzado en versión beta en noviembre de 2004, clasifica los resultados utilizando un algoritmo similar al que emplea Google para las búsquedas generales, aunque también tiene en cuenta la calidad de la revista en la que se ha publicado el artículo e incluye enlaces a otros artículos que lo citan. Permite buscar copias físicas o digitales de artículos, ya sea en línea o en bibliotecas. Para nosotros la modificación de la búsqueda avanzada permitió filtrar los resultados para mostrar únicamente los pertenecientes a una publicación o autor. Los resultados más relevantes para las palabras clave buscadas se listaron en primer lugar, según la clasificación del autor, el número de referencias que lo enlacen, y su importancia respecto de otra bibliografía académica, así como la clasificación de la propia publicación en la que aparecía el artículo. El módulo utilizado, (https://github.com/ckreibich/scholar.py/blob/master/scholar.py), se ha modificado para ajustarse al problema planteado y guardar los resultados obtenidos en una base de datos MySQL. A partir de ahí introducimos nuestro propio script de buscador interno y el externo lanzándolo a Google Scholar.

Para el buscador interno se empleó NLTK, NLTK (Natural Language Tool Kit) es la plataforma más utilizada en el entorno Python para trabajar con datos en lenguaje natural. Proporciona múltiples interfaces fáciles de usar, e incluye alrededor de 50 corpora y diversos recursos léxicos tales como WordNet, junto con un conjunto de bibliotecas de procesamiento de texto que permiten la clasificación, la tokenización, el stemming, el etiquetado, el análisis y el razonamiento semántico. La biblioteca NLTK permite que cualquier programa escrito en Python pueda invocar un amplio conjunto de algoritmos que sustentan las principales técnicas de procesamiento del lenguaje natural para la generación de métricas, frecuencia de términos, polaridad negativa/positiva de frases y textos, entre otras muchas técnicas. Junto con NLTK se suelen utilizar otros módulos útiles para procesar la información contenida en los documentos textuales, como *numpy* (para el tratamiento de matrices), *scipy* (para realizar cálculos científicos), *matplotlib* (para la realización de gráficos), etc. junto con estas herramientas impulsamos scikit-learn que es una herramienta muy eficiente para la minería y análisis de datos. Permite agrupar datos usando diversos algoritmos, como k-means, métodos aglomerativos, DBSCAN, agrupamiento jerárquico con ward, etc. También es posible clasificar objetos en categorías con SVM, algoritmo del vecino más próximo, "selvas aleatorias", etc. Disponía además de módulos para reducir la dimensionalidad, seleccionar el modelo más adecuado y extraer características de objetos, ya sean documentos textuales o imágenes.

Por ello el modelo SVM de scikit-learn nos permitió como entrada tanto vectores densos como dispersos, lo que facilitó su uso en corpus textuales grandes como los que manejábamos, en los que los documentos suelen estar representados por vectores de características con muchos valores iguales a 0. En nuestro procedimiento scikit-learn proporcionaba varios métodos para clasificación basados en SVM. SVC, NuSVC y LinearSVC son clases pueden realizar clasificaciones multi-clase en un conjunto de datos. SVC y NuSVC tienen métodos similares, pero aceptan conjuntos de parámetros algo distintos y tienen formulaciones matemáticas diferentes. LinearSVC es otra implementación para la clasificación de vectores de soporte para el caso de que el kernel sea lineal. Al igual que otros clasificadores, estos tres modelos toman como entrada dos arrays: un array *X* de tamaño [num_ejemplos, num_características] que contiene los ejemplos de entrenamiento, y un array *y* de etiquetas de clases, que pueden ser cadenas o enteros, cuyo tamaño es [num_ejemplos].

SVC y NuSVC implementan el enfoque "uno-contra-uno" (Knerr, Personnaz y Dreyfus, 1990) para la clasificación multi-clase. Sea *num_clases* el número de clases de la colección, estos modelos construyen *num_clases * (num_clases – 1) / 2* clasificadores, y cada uno entrena datos de dos clases distintas.

Por su parte, LinearSVC implementa la estrategia multi-clase "uno-contra-todos", que consiste en ajustar un clasificador por clase, es decir, se construyen *num_clases* clasificadores. Para cada uno de ellos, una clase se ajusta contra todas las demás.

Este enfoque es el más utilizado puesto que es más eficaz computacionalmente al requerir solo *num_clases* clasificadores, y además es más fácil de interpretar debido a que cada clase se representa por un solo clasificador, lo que favorece obtener conocimiento sobre la clase analizando su clasificador correspondiente.

Para esta primera fase se ha modificado un módulo de búsqueda avanzada de google scholar escrito en Python, *scholar.py*, para poder almacenar las URLs de los textos que se van a descargar en una base de datos MySQL. Posteriormente, utilizando la biblioteca de Python *urllib2*, que permite abrir una URL y guardar el documento enlazado en un almacenamiento local, se recopilan el total de documentos. Disponemos de un corpus de textos científicos escritos en inglés, francés y español por cada una de las líneas de investigación.

Los textos de cada uno de estos campos de investigación se almacenan en carpetas distintas, para después convertirlos a formato de texto simple para poder analizarlos. El motivo de seleccionar los artículos de diversos años es para evitar posibles sesgos. El módulo *scholar.py* emplea la biblioteca *Beautiful Soup* para analizar los resultados devueltos por el buscador, facilitando la obtención de los documentos en PDF publicados por los investigadores sobre los temas y disciplinas que nos interesaban.

1.-*Preprocesamiento de los documentos*. Para poder realizar el análisis de los textos, en primer lugar los convertimos en texto simple. El módulo *PDFMiner* permite realizar esta conversión con el método *TextConverter*, que lee cada uno de los artículos página a página. El texto extraído de esta forma se pasa al módulo de *NLTK* para eliminar las *stopwords*, los signos de puntuación y los números, así como para normalizar los términos utilizando un lematizador de *WordNet* integrado en el módulo. Se extraen de esta forma las raíces de las palabras según se trate de un nombre, verbo o adjetivo. También se escriben todas las palabras de los documentos en minúscula.

Todo este pre-procesamiento del texto es útil en el trabajo que estamos llevando a cabo porque solo nos interesan las palabras, no los números o signos de puntuación, ni tampoco nos interesa en esta fase el reconocimiento de entidades nombradas, por eso se escribe todo en minúscula.

Se realiza esta normalización documento a documento, guardando el resultado en su correspondiente carpeta.

2.-*Extracción de características.* Las características que definen un documento textual son los n-gramas más utilizados en el mismo. Se suelen extraer unigramas, bigramas y trigramas, ya que los términos formados por más de tres palabras pueden dar peores resultados y añadir ruido a la correcta categorización textual. Python dispone de un módulo denominado *collections* especializado en la generación de contenedores en forma de diccionarios, listas y conjuntos. Los diccionarios generados con *Counter* contienen pares *atributo-valor* que incluyen los términos de un documento y el número de veces que aparece cada uno de ellos. Puesto que uno de nuestros objetivos es la construcción de un vocabulario que defina cada uno de los grupos de la colección de documentos descargados, extrajimos todos los unigramas, bigramas y trigramas que aparecieron en cada grupo, así como el número de veces que aparecieron en el total de los documentos. Aunque se han eliminado los que aparecen solo una vez (normalmente son errores ortográficos o términos tan infrecuentes que no aportan información relevante) el número de n-gramas obtenido de esta forma es muy elevado.

Selección de características

La representación de los documentos como unigramas, bigramas y trigramas es el enfoque más simple del procesamiento del lenguaje natural. Se denomina "bolsa de palabras" (bag of words), y asigna un peso a cada término en función de su importancia, determinada por su frecuencia de aparición en el documento.

Este método suele dar resultados aceptables para una primera aproximación, pero es mejor utilizar otros que, aunque son más complejos, son más útiles.

Uno de los más utilizados es TF-IDF (Term Frequency, Inverse Document Frecuency). Se trata de un método heurístico que calcula la importancia de un término según su frecuencia de aparición en el documento pero teniendo en cuenta su frecuencia de aparición en el conjunto total de documentos de la colección.

Actualmente se utiliza un método probabilístico basado en la divergencia entre los documentos de la colección con el que se obtienen muy buenos resultados. Se denomina Kullback-Leibler Divergence (KLD), y consiste en calcular el peso de un término como: $KLD = pD(t) \cdot ln\left(\frac{pD(t)}{pC(t)}\right)$

Donde *pD(t)* es la probabilidad de cada término *t* en el documento *D* y *pC(t)* es la probabilidad del mismo término *t* en toda la colección. En nuestro caso queremos calcular la probabilidad de los términos en el conjunto de los 375 documentos de cada grupo en relación con toda la colección.

De esta forma se asigna un peso mayor a aquellos términos que aparecen con frecuencia en un grupo y no aparecen o aparecen con poca frecuencia en los demás grupos que forman la colección.

El objetivo es seleccionar los términos que son específicos de cada grupo, descartando o asignando un peso menor a los que aparecen en la mayoría de los documentos de la colección.

La correcta selección de términos es fundamental para la representación del texto y para conseguir mejores resultados del sistema de aprendizaje. Uno de los principales problemas en el procesamiento del lenguaje natural es la alta dimensionalidad tanto en el número de ejemplos (documentos a analizar) como en el número de características (unigramas, bigramas, etc., que definen cada documento). Cada documento se representa como un vector de características, utilizando para ello la biblioteca *numpy* de Python, que permite además realizar diversos cálculos científicos. Tendremos un vector de dimensión (*n_samples x n_features*). En este trabajo tenemos 6000 ejemplos, por lo que para no tener problemas de memoria

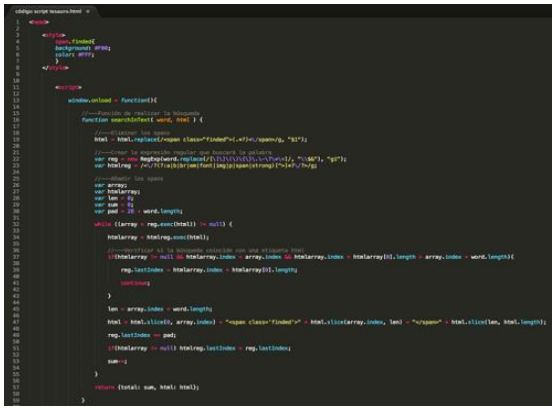
al representar cada uno de ellos como un vector de características (unigramas, bigramas y trigramas) debemos hacer una selección de las mismas antes de utilizar algún método de agrupamiento o clasificación de los documentos.

Existen diversos modelos para reducir la dimensión de las características, como *VarianceThreshold*, un enfoque muy simple que elimina aquellas características (unigramas, bigramas o trigramas en nuestro caso) cuya varianza no llega a un límite establecido, X^2*-test*, o incluso la *fuerza del término*. También se pueden usar algunos de los modelos disponibles en la biblioteca *gensim* de Python, como *lda* (Latent Direchlet Allocation), *word2vec*, *lsi* (Latent Semantic Indexing) o *tf-idf*.

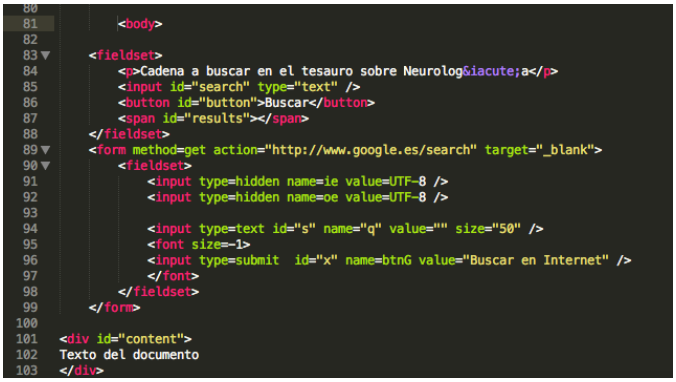
Una de las principales propiedades de la representación textual como un vector de características es que los vectores son dispersos, es decir, muchos de su valores son 0 (la mayoría de los términos no aparecen en todos los documentos). Esto permite usar matrices o vectores dispersos, cuyas operaciones son mucho más rápidas y eficaces que con vectores densos. Lo cual nos da un valor de representatividad elevado.

6. Script del buscador Glosarios (Head) 1

7. Script del buscador Glosarios (Body) 2



8. Script del buscador Tesauros (Head) 1



9.Script del buscador Tesauros (Head) 2

10. Script del buscador Tesauros (Body) 3

Recibidos los trabajos, se han evaluado por los profesores tutores, en el caso de los alumnos de TFG, y por el resto de los profesores en el resto[3].

4. Resultados obtenidos y Discusión

Como hemos podido ver a lo largo de este trabajo, podríamos decir que el proyecto ha constado de dos vertientes: por un lado, la vertiente de creación de herramientas desde el punto de vista del contenido y, por otro, la creación de herramientas desde el punto de vista técnico.

[3] Algunos de los trabajos, debido a la gran cantidad de términos que poseen y a la respuesta masiva por parte del alumnado de la asignatura de Traducción de Textos Científico-Técnicos, aún están en evaluación (más de 900 términos por glosario e idioma).

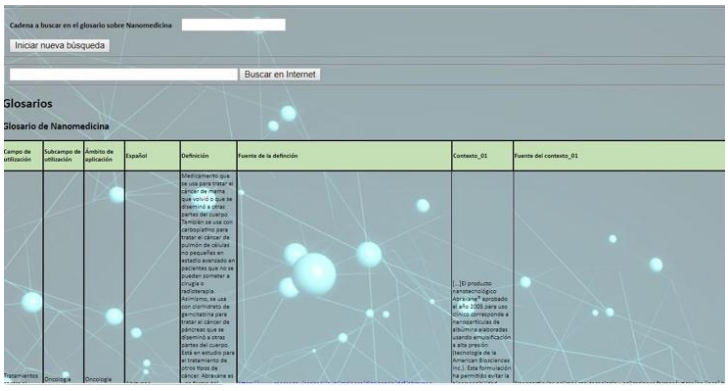
Una vez creado, evaluado y organizado todo el contenido (glosarios y tesauros), se ha tenido que realizar el volcado en el sitio web llamado UCOTerm: Laboratorio de Ingeniería Documental, un sitio web ubicado en la Universidad de Córdoba que ha nacido a partir de un primer proyecto de innovación docente «UCOTerm: Sitio web para la difusión de recursos para la traducción científico-técnica» , pero que pretende ir mucho más allá y cubrir otros ámbitos temáticos de interés para traductores e investigadores.

11. Vista del inicio de UCOTerm

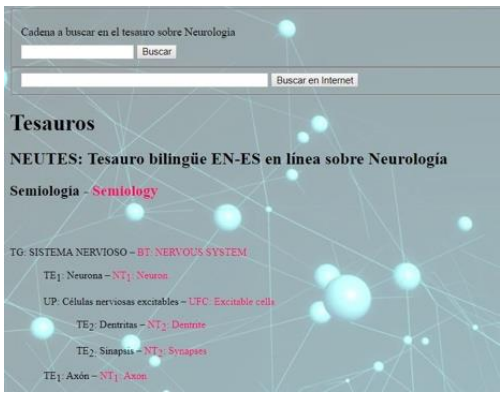
UcoTerm tiene una interfaz sencilla que permite al usuario navegar por ella fácilmente y encontrar las herramientas documentales que más le interesan. Como se puede apreciar en la figura 12, los glosarios de carácter científico-técnico aparecen de manera alfabética tras seleccionar la pestaña «GLOSARIOS»:

12. Vista de la pestaña GLOSARIOS

Actualmente dispone de 7 glosarios especializados con autoría de alumnos y supervisión del profesorado, lo que ha supuesto un alto grado motivador para alumnado y profesorado. Asímismo, contamos con cuatro glosarios más en evaluación, algunos de ellos con más de 900 términos que espereamos poder tener publicados para finales de este curso. Para consultar los glosarios, únicamente hay que entrar en el que nos interesa e introducir nuestra búsqueda en «Cadena a buscar en el glosario» si deseamos buscar en el glosario o en «buscar en Internet» si además deseamos ampliar nuestra búsqueda a toda la red (*cf.* Figura 13).



13. Ejemplo de glosario



14. Ejemplo de tesauro

La pestaña «TESAUROS» (*cf.* Figura 14) no dista mucho de la de los «GLOSARIOS». De momento hay uno publicado y cuatro en evaluación y normalización.

El sistema de búsqueda en los tesauros es similar a la de los glosarios, por lo que resulta muy intuitiva para el usuario, pudiendo mostrar una jerarquía de términos normalizada al usuario junto con sus equivalentes.

5. CONCLUSIONES

Gracias a este proyecto, se ha puesto en marcha un sitio web que diufunde y difundirá las herramientas y recursos de calidad superior desarrolladas por el alumnado de las asignaturas de Documentación, Herramientas Profesionales para la Traducción, Traducción Científico-Técnica, Trabajo de Fin de Grado y de Máster, respetando su autoría.

Asimismo, se ha puesto de maniefiesto que el alumnado que ha colaborado en estre proyecto se ha sentido altamente motivado, lo que ha repercutido directamente en el aprendizaje de éste y en la calidad de los trabajos publicados. Por otro lado, se ha fomentado el autoaprendizaje como medio de desarrollo, innovación y responsabilidad profesional a través de la formación práctica, continua y especializada. El alumnado ha aprendido a investigar, a contrastar, a traducir y a trabajar acorde unos plazos fijados con unos resultados muy satisfactorios por parte del profesorado.

Igualmente, se ha potenciado la competencia instrumental del alumnado gracias a la gestión y creación de recursos de documentación para la resolución de problemas de traducción, intengrado las TIC como herramientas que contribuyen a que el estudiante abandone el papel de sujeto receptor y pase a ser un elemento activo de su aprendizaje.

Del mismo modo, gracias a UCOTerm, se facilita la comunicación multilingüe entre profesionales e investigadores del ámbito científico-técnico, lo cual es enriquecedor tanto para el alumnado, como para el profesorado que hemos colaborado en el proyecto.

Por último, con UCOTerm ofrecemos un repositorio de recursos documentales esenciales y requeridos para la práctica profesional de la traducción especializada que esperamos siga creciendo en esta línea estableciendo lazos colaborativos entre alumnos y profesores.

BIBLIOFRAFÍA

BIBLIOTECA DE RECURSOS PARA BIBLIOTECARIOS Y OPOSITORES (2017). «Lenguajes documentales. Los tesauros: creación y mantenimiento. NORMA ISO 2788». *Bibliopos.es. Biblioteca de Recursos para Bibliotecarios y Opositores*. Disponible en: http://www.bibliopos.es/Biblion-A2-Bibliografia-Documentacion/11Lenguajes-documentales-Tesauros-ISO-2788.pdf [Consulta: 13/04/2018]

CABRÉ, M. T., Y CASTELLVÍ, M. T. C. *La terminología: teoría, metodología, aplicaciones*. Antártida/Empúries. 1993.

CASTILLO PEREIRA, I. *Acerca del lenguaje científico-técnico. Sus características y clasificación.* Departamento de TRADUCCIONES CNICM-Informed. DIT. 2015.

CORPAS PASTOR, G. «Compilación de un corpus "ad hoc" para la enseñanza de la traducción inversa especializada». *TRANS*. n.º 5, 155-184. 2001.

CORPAS PASTOR, G., Y SEGHIRI DOMÍNGUEZ, M. «Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor». *Procesamiento del Lenguaje Natural,* 39, 165-172. 2007.

FERNÁNDEZ-ALTUNA, M.A. *ET AL. Uso de los MeSH: una guía práctica*. Inv. Ed. Med. 5 (20):220-229. 2016.

FINKEL, J.R., GRENAGER, T., AND MANNING, C.. «Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling». *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* pp. 363-370. 2005. URL: http://nlp.stanford.edu/software/CRF-NER.shtml

HERRERO, C. M. «Aproximación a ciertas perspectivas en la lingüística de corpus». En: A. GÓMEZ, y A. SÁNCHEZ Pérez A *Survey en Corpus-based Research/Panorama de investigaciones basadas en corpus*. Murcia: AELINCO. 2009.

KNERR, S.; PERSONNAZ, L.; DREYFUS, G.. «Single-layer learning revisited: a stepwise procedure for building and training a neural network». En *Neurocomputing*. Springer Berlin Heidelberg, 1990. pp. 41-50.

LEÓN, P. *El Traianeum de Italica*, Sevilla, 1988.

MANNING, Christopher D.; SCHÜTZE, Hinrich. *Foundations of statistical natural language processing*. Cambridge: MIT press, 1999.

MOGOLLÓN, G. «Paradigma científico y lenguaje especializado». *Revista de la Facultad de Ingeniería de la Universidad Central de Venezuela*, 18(3), 5-14. 2003.

MURPHY, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

PARODI, G. «Lingüística de corpus: Una introducción al ámbito». *RLA. Revista de Lingüística Teórica y Aplicada*. 46 (1), I Sem. 2008, pp. 93-119. 2008.

SALTON, G.; WONG, A.; YANG, C. «A vector space model for automatic indexing». *Communications of the ACM*, 1975, vol. 18, no 11, pp. 613-620.

SUTTON, Richard S.; BARTO, Andrew G. *Reinforcement learning: An introduction*. MIT press, 1998.

TREUHERZ, A.; RIBERIO, O. *DeCS – Descriptores en Ciencias de la Salud*. 2007.

UNIVERSIDAD DE SALAMANCA (2017) «Tesauros: conceptos, elaboración y mantenimiento». *Biblioteca de Traducción y Documentación.* Disponible en: http://sabus.usal.es/docu/pdf/Tesauro.PDF [Consulta: 16/04/2018]

VARGAS SIERRA, C. (2006). Diseño de un corpus especializado con fines terminográficos: el corpus de la piedra natural.