http://dx.doi.org/10.1590/1678-460X2019350405

Metodología para la construcción de descriptores en la recuperación de información aplicada a la traducción especializada (EN/FR/ES): el caso de vehículos aéreos no tripulados

Metodologia para a construção de descritores na recuperação de informações aplicadas à tradução especializada (EN/FR/ES) sobre veículos aéreos não tripulados.

A RI methodology to build descriptors for specialized translation (EN/FR/ES) about Air Vehicles Unmanned.

María Azahara Veroz González[1]

RESUMEN

En este trabajo desarrollamos un marco metodológico para la construcción empírica y no intuitiva de una sintaxis de búsqueda para recuperar información dentro del proyecto OntoUAV: "Ontología web multilingüe (EN / FR / IT / ES) sobre vehículos aéreos no tripulados para la traducción de textos" en el ámbito de los vehículos aéreos no tripulados.

1. Profesora ayudante doctor en el Departamento de Traducción e Interpretación, Filología Francesa, Estudios Semíticos y Documentación. Universidad de Córdoba – España. http://orcid.org/0000-0001-9544-4090. E-mail: z92vegom@uco.es.

OntoUAV es un proyecto terminológico y ontológico que dará respuesta a las necesidades lingüísticas actuales en esta materia específica, teniendo en cuenta el nuevo marco normativo que debe transponerse a los Estados miembros de la UE. En primer lugar, para llevar a cabo este trabajo, desarrollamos una metodología para la extracción y construcción de una sintaxis de búsqueda que podría ser utilizada para compilar un corpus de textos para la traducción especializada en este ámbito o para la futura creación de herramientas terminológicas y/o ontológicas.
Palabras clave: *recuperación de información; recuperación de información para traducción; descriptores; Ontología; traducción especializada; herramientas profesionales para traductores; drones.*

ABSTRACT

In this work, we develop a methodological framework for the empirical and non-intuitive construction of a search syntax in other to retrieve information either within the OntoUAV project: "Multilingual web ontology (EN / FR / IT / ES) on Air Vehicles Unmanned for texts translation" or within specialized translation in the field of unmanned aerial vehicles. OntoUAV is a terminological and ontological project that will answer the current linguistic needs in this specific subject, considering the new regulatory framework that must be transposed to the EU Member States. First, to carry out this work, we developed a methodology for the extraction and construction of a search syntax, that could be used to compile a corpus of texts for specialized translation in this field or for the future creation of terminological and / or ontological tools.
Keywords: *information retrieval; information retrieval for translation; descriptors; Ontology; specialised translation; professional tools for translators; drones.*

RESUMO

O presente trabalho aborda uma metodologia para a construção empírica (e não intuitiva) de uma sintaxe que abarca a procura apropriada da recuperação de informação relacionada ao projeto OntoUAV "Ontologia web multilíngue (EN/FR/IT/ES) de Veículos Aéreos não Tripulados orientada à tradução de textos", além da tradução especializada no âmbito dos veículos aéreos não tripulados. OntoaUAV é um projeto terminológico e ontológico que visa atender às necessidades lingüísticas atuais nesse sentido específico, considerando o novo enquadramento regulatório da UE o qual se deve transpor às regulamentações dos estados membros. A fim de

DELTA
35.4
2019

realizar o trabalho deve-se, primeiramente, desenvolver uma metodologia de extração e construção de uma sintaxe de investigação que possa ser utilizada na compilação de um corpus de textos de especialidade para a tradução nesse terreno, assim como no que tange a criação futura de ferramentas terminológicas ou ontológicas.
Palavras-chave: *recuperação de informação; documentação aplicada à tradução; descritores; ontologia; tradução especializada; ferramentas profissionais para a tradução; drones.*

1. Introducción

En el siguiente trabajo se expone la metodología desarrollada para la construcción de descriptores en la recuperación de información aplicada a la traducción especializada (EN/FR/ES), aplicándose al caso de vehículos aéreos no tripulados. Esta metodología se ha desarrollado en el seno del proyecto OntoUAV: "Ontología web multilingüe (EN/FR/IT/ES) sobre Vehículos Aéreos no Tripulados orientada a la traducción de textos" a fin de responder a las necesidades lingüísticas creadas por la Directiva 2009/48/CE y el Reglamento (CE) No 216/2008, que han de trasponerse a los Estados miembros de la Unión Europea.

En el diseño de este proyecto, advertimos que, en el ámbito de la documentación no existía una propuesta metodológica completa a la hora de establecer los descriptores para la creación de una sintaxis de búsqueda en la recuperación de información en ámbitos especializados. Por ello, nos propusimos diseñar esta metodología de forma protocolizada, con el fin de responder no solo a las carencias que debíamos suplir en este proyecto, sino también en la traducción especializada en general.

1.1. Contexto europeo

El Reglamento (CE) Nº 216/2008 exige a la Agencia Europea de Seguridad Aérea (EASA) regular los Vehículos Aéreos no Tripulados (VANT), en concreto, aquellos vehículos de uso civil y cuya masa sea superior a 150 kg.

El ámbito de aplicación de este Reglamento se extiende, según su artículo 1:

> a) Al diseño, producción, mantenimiento y operación de productos, componentes y equipos aeronáuticos, así como al personal y organizaciones que participen en el diseño, la producción y el mantenimiento de tales productos, componentes y equipos aeronáuticos.
> b) Al personal y organizaciones que participen en la explotación de aeronaves.
>
> Artículo 1 del Reglamento (CE) Nº 216/2008

Este Reglamento excluye a aquellos que formen parte de operaciones militares, aduaneras, de policía o similares, y a vehículos aéreos no tripulados cuyo peso sea igual o inferior a 150 kg. Estos últimos quedan regulados por la Directiva 2009/48/CE sobre la seguridad de los juguetes, ya que los drones de este peso serán considerados juguetes, mientras que aquellos que tengan un peso superior a 150 kg serán considerados aeronaves.

Estas normativas, por el principio de supremacía de la Unión Europea, obligan a los Estados miembros a adaptar sus ordenamientos jurídicos para que tengan en cuenta lo dispuesto en el Reglamento y en la Directiva.

Según esto, muchos Estados miembros han reformado ya sus ordenamientos jurídicos[2] para aplicar tales normativas, otros están en vías y otros, como es el caso de España, tienen normativas provisionales a la espera de que se aprueben leyes definitivas. En cualquiera de los casos, nos encontramos ante un mercado prácticamente virgen, no solo en el ámbito jurídico, en donde será necesaria la creación de recursos que faciliten la trasposición de nuevas estructuras jurídicas, sino también en el ámbito técnico. En Europa nos encontramos con un mercado emergente, siendo el líder en la fabricación de estos sistemas el estadounidense y el canadiense, de hecho, tras realizar varias búsquedas, son pocos los proveedores hispanohablantes, por lo que es necesario disponer de estrategias documentales científicas y recursos lingüísticos basados en esas estrategias documentales para la traducción

2. Austria, República Checa, Dinamarca, Francia, Alemania, Irlanda, Italia, Países Bajos, Suecia, Suiza, Reino Unido. Fuente: EASA.

de los textos, que actualmente se están produciendo en otros idiomas como es el inglés y el francés.

1.2. La recuperación de información en el ámbito de la Documentación Aplicada a la Traducción

La Documentación Aplicada a la Traducción es una disciplina y una práctica indispensable dentro del proceso traductor, ya que el conocimiento intralingüístico ha de completarse con el extralingüístico y especializado y, para hacer frente a estas dos últimas competencias, el conocimiento extralingüístico y el especializado, el traductor debe desarrollar una buena estrategia documental.

Esta estrategia documental se podrá desarrollar recuperando dos conceptos básicos de las Ciencias de la Documentación: la recuperación de información (en inglés, *information retrieval*) que a partir de ahora denominaremos RI y los sistemas de recuperación de información (SRI).

Consideramos fundamentales las definiciones planteadas por Croft (1987) y Korfhage (1997) sobre RI, ya que plantean explícitamente el papel del usuario como fuente de las consultas y destinatario de las respuestas; en este caso, el usuario sería el traductor especializado pues es el que reclama la información para su traducción, en esta línea, la RI se definiría como:

> El conjunto de tareas mediante las cuales el usuario localiza y accede a los recursos de información que son pertinentes para la resolución del problema planteado. En estas tareas desempeñan un papel fundamental los lenguajes documentales, las técnicas de resumen, la descripción del objeto documental, etc. (Croft 1987).
>
> La localización y presentación a un usuario de información relevante a una necesidad de información expresada como una pregunta. (Korfhage 1997).

Es decir, se podría afirmar que la RI tiene como objetivo encontrar y diferenciar aquellos documentos, imágenes, vídeos, etc., que satisfagan las necesidades del usuario, expresadas por él mismo. Pero esto puede plantear al usuario varios problemas:

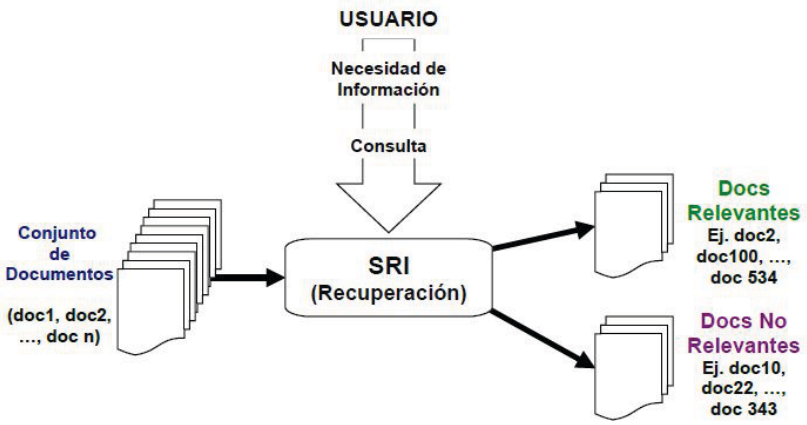
– Cómo compatibilizar y comparar el lenguaje en que está expresada tal necesidad de información y el lenguaje de los documentos.

– Y, la que nos ocupa principalmente en este trabajo: cómo plantear la consulta apropiada.

La primera cuestión podría resolverse por la propuesta realizada por Baeza-Yates (1999), que abarca el problema desde dos perspectivas: el computacional y el humano. El computacional, relacionado con la construcción de estructuras de datos y algoritmos eficientes que mejoren la calidad de las respuestas y el humano, con el estudio del comportamiento y de las necesidades de los usuarios.

Tolosa y Bordignon (2011) lo explican de la siguiente manera:

Ilustración 1 – Interacción entre el usuario y los SRI (Tolosa y Bordignon, 2011)



– Existe una colección de documentos que contienen información de interés (sobre uno o varios temas).
– Existen usuarios con necesidades de información, quienes las plantean al SRI en forma de una consulta (en inglés, *query*).
– Como respuesta, el sistema retorna – de forma ideal – referencias a documentos "relevantes", es decir, aquellos que satisfacen la necesidad expresada, generalmente en forma de una lista ordenada.

En la segunda cuestión, *cómo plantear la consulta apropiada,* hemos de desarrollar una serie de estrategias que nos ayuden a mejorar y aumentar la funcionalidad de los SRI, como son:

- *La extracción de información:* Se consideraría como aquella metodología usada para extraer aquellas porciones de texto que poseen alta carga semántica y establecer relaciones entre los términos. Es lo que otros, Jacquemin y Tzoukermann (1999), llaman generación de términos índice o normalización (*conflation*). Esta metodología es clave para el trabajo que aquí proponemos.

- *Los modelos de recuperación:* se trata de aquella metodología utilizada para establecer cómo representar los documentos y las necesidades del usuario. Es decir, se plantean desde el punto de vista del diseño del SRI. No obstante, es importante que el usuario-traductor conozca cuáles son los principales modelos existentes y sobre cuáles se basan sus principales buscadores para así diseñar una sintaxis de búsqueda apropiada.

- *Búsquedas web:* se refiere a los SRI que operan en un corpus web, ya sea privado (intranet) como público (internet). Internet ha planteado nuevos desafíos debido a sus características propias: dinámica, tamaño, datos no estructurados. En el tema que aquí nos ocupa, nos dedicaremos a cómo extraer información fidedigna de Internet.

La recuperación en línea en redes y los sistemas que la facilitan en Internet poseen una gran movilidad y adaptabilidad en el tratamiento funcional de los datos y en la variedad y la multiplicidad de productores que determinan una gran redundancia y, a la vez, una dispersión de la información distribuida por fuentes y respuestas; cada buscador, cada sistema de recuperación, indiza solo una parte del universo informacional que contiene Internet. Además, existe una gran diversidad de soportes que causan una gran heterogeneidad en la estructura de los documentos a recuperar (videos, gráficos, multimedia, etc.), no olvidando que mucha información en Internet está caracterizada por el dinamismo y la volatilidad. Este dinamismo se refiere a los continuos cambios de contenido de muchos de los documentos de Internet y la volatilidad, a los cambios de destino de un mismo docu-

mento. A esto hay que sumar que la investigación sobre la recuperación de información y su análisis tiene una amplia repercusión en el campo de los lenguajes controlados, es obvio si pensamos que ese es uno de sus principales objetivos: el control de la información a través de un sistema universal.

El ámbito de la traducción especializada no ha tenido tanta suerte, aunque su importancia en el proceso informacional hoy día es crucial frente a la expansión y desarrollo de la información en los nuevos entornos. Todo tipo de código y mensaje se ha editado a lo largo del tiempo para la comprensión de la información encauzado por cierto tipo de especialistas que las sociedades poseen y que dedican sus esfuerzos al nacimiento, desarrollo, mantenimiento y difusión de la información, ya sea desde un plano teórico, práctico o mítico. Tras el proceso de globalización y su difusión, los procesos de producción y difusión de la información se han multiplicado y, por este motivo, es necesario introducir herramientas que auxilien a las ya existentes en el proceso de control, tanto a los profesionales como a los neófitos. Uno de los aspectos que se tratan en el proceso de tratamiento de los datos para su posterior difusión es su presentación gráfica, por ello, en el marco de creación de un sistema de creación de recuperación documental basado en descriptores sintagmáticos, con enlaces de acceso y captura a ontologías OWL, son necesarios ciertos elementos, tales como el análisis documental y el rigor técnico. Es importante no prescindir de ninguno de estos dos elementos, ya que se corre el riesgo de mutilar el trabajo de recuperación. Al igual que en muchos otros ámbitos del conocimiento, los sistemas de recuperación poseen un problema de sinonimia y control de descriptores, lo que lleva a destacar la importancia y la necesidad de un lenguaje controlado que permita un tratamiento sistemático del vocabulario con vistas a la indización y la recuperación terminológica en este ámbito, y al análisis de la metodología de la calidad informativa lo cual es el objeto del presente trabajo.

Este texto pretende, por tanto, ayudar también a los traductores en esta labor de comprensión en la medida de sus posibilidades, con ello se mejora la calidad informativa y, por ende, la calidad del producto de información traductológica como producto global, que puede emplearse tanto en el proceso teórico de información, como en el prác-

tico de documentación, así como también en el teórico-práctico de la localización traductológica.

Sin embargo, a pesar de la heterogeneidad de la red existen algunos patrones generales tanto en el comportamiento de los usuarios como en la forma en que se crea y comparte información. De este modo, nos encontramos que coexisten sitios compuestos de miles de páginas, con miles de sitios que contienen sólo unas pocas páginas. Esta proporción y relación directa se hereda en el comportamiento de la información enlazada: pocos sitios contienen miles de enlaces, pero muchos sitios contienen dos o tres, y, a su vez, este comportamiento de la información es compartido con los usuarios: millones de usuarios se unen en unos pocos sitios preferidos, prestando poca atención a otros millones de sitios. Estas relaciones podrían considerarse como un punto de inflexión paradigmática en los comportamientos científicos, culturales y sociales actuales, y un estudio en profundidad de sus procesos sería ideal para una comprensión objetiva de la evolución de las sociedades informativas actuales (Marcos Aldón 2016).

1.3. Definición de DTD y el XML Schema

En los últimos años, nos encontramos ante la necesidad de estructurar y jerarquizar la información para almacenarla y facilitar su acceso y recuperación a través de los SRI adecuados. Por este motivo, se crearon las DTD (*Document Type Definition*, Definición de Tipo de Documento) y los Esquemas XML (XML Schema).

Una DTD es un documento que define la estructura de un documento XML (*eXtensible Markup Language* o Lenguaje de Marcado eXtensile) o SGML (*Standard Generalized Markup Language* o Lenguaje de Marcado Generalizado Estándar). Las DTD incluyen una serie de reglas sintácticas para un tipo de documento específico, es decir, incluyen los elementos que se permiten y sus atributos, así como también las reglas que afectan a la anidación de los primeros y a los valores de los segundos.

Las DTD cumplen las siguientes funciones:

- Especificar las clases de documentos:
  - Describiendo un formato de datos.
  - Utilizando un formato común de datos entre aplicaciones.
  - Verificando los datos al intercambiarlos.
  - Verificando un mismo conjunto de datos.
- Describir:
  - Elementos: cuáles son las etiquetas permitidas y cuál es el contenido de cada etiqueta.
  - Estructura: en qué orden van las etiquetas en el documento.
  - Anidamiento: qué etiquetas van dentro de cuáles. (Lamarca, 2013).

Y se componen de elementos y atributos, especificando cómo se anidan ente ellos:

Los elementos según Lamarca (2013) son los siguientes:

- Elementos con "contenido ELEMENT":
  - Un elemento tiene contenido ELEMENT, si solo puede contener a otros elementos, opcionalmente separados por espacios en blanco.
- Elementos con "contenido TEXT":
  - Un elemento tiene contenido TEXT, si solo puede contener texto.
  - (PCDATA = printable character data)
- Elementos con "contenido MIXED"
  - Un elemento tiene contenido MIXED, si puede contener texto u otros elementos
- Elementos con "contenido EMPTY"
  - Un elemento tiene contenido EMPTY, si no puede contener otros elementos

DELTA
35.4
2019

Y los atributos son los siguientes:

• CDATA: texto
• NMTOKEN: “abc...z0123..9-_:.” (tipo de lista)
• NMTOKENS: NMTOKEN + espacios
• ID: empezar con letra
• IDREF: ser un ID. (Lamarca, 2013).

A través del procesador XML, el navegador puede verificar si un documento es válido a través de la DTD, es decir, si el documento cumple las reglas del DTD. En este caso, la DTD se indica mediante la etiqueta DOCTYPE que, puede escribirse tanto internamente del documento XML como externamente, en cuyo caso se guardaría en un archivo de texto como, por ejemplo, en el archivo «drones.dtd». Todos los documentos DTD, han de tener un elemento raíz, siendo éste el que debe aparecer después de DOCTYPE, en el ejemplo descrito sería “drones”.

Por ejemplo, si la DTD fuera en un documento externo:

Ilustración 2 – DTD en un documento externo (I)

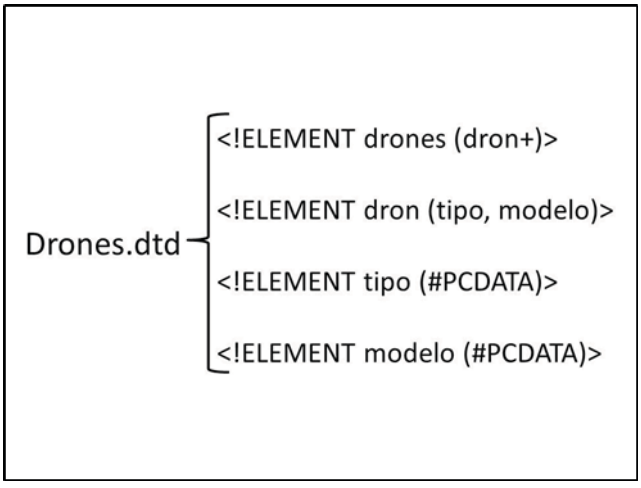


Ilustración 3 – DTD en un documento externo (II)

```
<?XML version="1.0 enconding="utf-8"?>
<!DOCTYPE Drones SYSTEM "Drones.dtd">
<Drones>
        <dron>
                <tipo>cuatrimotor</tipo>
                <modelo>Parrot Bepop 2</modelo>
        </dron>
        <dron>
                <tipo>cuatrimotor</tipo>
                <modelo>Parrot AR.Drone</modelo>
        </dron>
</Drones>
```

Y si fuera en el mismo documento:

Ilustración 4 – DTD en el mismo documento

```
<?XML version="1.0 enconding="utf-8"?>
<!DOCTYPE Drones [
                <!ELEMENT drones (dron+)>
                <!ELEMENT dron (tipo, modelo)>
                <!ELEMENT tipo (#PCDATA)>
                <!ELEMENT modelo (#PCDATA)>
                ]>
<Drones>
        <dron>
                <tipo>cuatrimotor</tipo>
                <modelo>Parrot Bepop 2</modelo>
        </dron>
        <dron>
                <tipo>cuatrimotor</tipo>
                <modelo>Parrot AR.Drone</modelo>
        </dron>
</Drones>
```

Cierto es que cuando un navegador accede a un documento puede comprobar a través de la información contenida en la DTD si el documento es válido o no y, según esto, decidir cómo tratarlo. Sin embargo, en la práctica, aunque el documento sea inválido, el navegador lo

interpreta, aunque sin la DTD es posible que no lo haga de la manera adecuada. Por ello, es necesario que la DTD esté bien definida, para que el documento se trate y se organice correctamente.

No obstante, las DTD indican solo qué elementos y atributos contiene un documento y cómo se anidan, pero no hace referencia a los tipos de datos (excepto el CDATA de texto plano), por eso, se utiliza también el Esquema XML (*Schema XML*), ya que suple esta carencia.

Como hemos comentado anteriormente, los Esquemas o *Schemas XML* describen la estructura y el contenido de la información presente en un archivo XML, tal y como lo hacían las DTD, pero de una manera más detallada, ya que además de elementos y atributos contienen tipos de datos, número máximo y mínimo de ocurrencias y otro tipo de datos más específicos.

De acuerdo con la especificación del *W3C XML Schema* (http://www.w3.org/XML/Schema), los esquemas expresan vocabularios compartidos que permiten a las máquinas extraer las reglas que han sido realizadas por las personas. Definiendo la estructura, contenido y semántica de los documentos XML.

Las ventajas de utilizar los esquemas son las siguientes:

- Emplean sintaxis de XML, al contrario que las DTD.
- Permiten especificar los tipos de datos.
- Son extensibles (es decir, permiten crear nuevos elementos).
- Son más ricos semánticamente.
- Siguen la sintaxis estándar de XML.
- Soportan *namespaces*.
- El *XML Schema* es un estándar de W3C. (Fraga, 2010; Lamarca, 2013).

Según esto, a diferencia de las DTD, el esquema nos permitiría definir los tipos de contenido de los elementos o atributos, especificando si van a ser fechas, cadenas de texto, números, etc.

Un *XML Schema* define:

- Los elementos que pueden aparecer en los documentos
- Los atributos que pueden aparecer en un documento

- Qué elementos son hijos
- El orden de los hijos
- El número de elementos hijos
- Si un elemento está vacío o puede incluir texto.
- Los tipos de datos de sus elementos y atributos
- Los valores por defecto y fijos para elementos y atributos (Fraga 2010).

La característica más reveladora del *XML Schema* es que soporta diferentes tipos de datos, permitiendo describir los elementos autorizados si los datos son correctos (Fraga 2010).

Por otro lado, tal y como sucedía con las DTD, el *XML Schema* ha de estar bien formado, y cumplir con las siguientes reglas sintácticas de XML:

- Empezar con la declaración XML.
- Tener un único elemento raíz.
- Las etiquetas de apertura deben tener sus correspondientes etiquetas de cierre.
- Los elementos son "case-sensitive".
- Todos los elementos deben cerrarse.
- Todos los elementos tienen que estar adecuadamente anidados.
- Los atributos deben estar entre comillas.
- Se deben utilizar las entidades para utilizar caracteres especiales. (Fraga, 2010)

Por ejemplo, partiendo del documento XML:

Ilustración 5 – Documento de partida oara el XML Schema

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<drones>
<dron>
<nombre>Parrot Bepop 2</nombre>
<tipo>cuadrimotor</tipo>
<fabricacion> <inicio> <dia>21</dia> <mes>julio
</mes> <anyo>2014</anyo> </inicio>
 <fin> <dia>9</dia> <mes>August</mes> <anyo
>2014</anyo> </fin> </fabricacion>
 </dron> </drones>
```

El *XML Schema* sería:

Ilustración 6 – XML Schema

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<vehiculos  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="XMLSchemaBasicos02_drones.xsd">
 <dron>
  <nombre>Parrot Bepop 2</nombre>
  <tipo>cuadrimotor</tipo>
  <fabricacion>
   <inicio>
        <dia>21</dia>
        <mes>julio</mes>
        <anyo>2014</anyo>
   </inicio>
   <fin>
        <dia>9</dia>
        <mes>agosto</mes>
        <anyo>2014</anyo>
   </fin>
  </fabricacion>
 </vehiculo>
</vehiculos>
```

El W3C establece una serie de recomendaciones con respecto a los *XML Schemas* y que especifica Lamarca (2013):

- XML Schema Part 0: Primer. Es un documento no normativo que pretende ofrecer una fácil descripción de las funcionalidades de XML Schema y que está orientado a comprender de forma rápida cómo crear esquemas utilizando el lenguaje XML Schema http://www.w3.org/TR/xmlschema-0/.
- XML Schema Part 1: Structures. Especifica la definición del lenguaje XML Schema y ofrece las herramientas para describir la estructura y constreñir el contenido de los documentos de XML 1.0, incluyendo las que tratan los espacios de nombre (XML Namespace). El lenguaje de esquemas que se representa en XML usa espacios de nombre, reconstruye sustancialmente y extiende las capacidades de las DTDs de los documentos del lenguaje XML 1.0 http://www.w3.org/TR/xmlschema-1/.

- XML Schema Part 2: Datatypes. Establece las herramientas para definir los tipos de datos que se usan en los esquemas XML y en otras especificaciones XML. El lenguaje de tipo de datos que se representa en XML 1.0, ofrece un conjunto de capacidades que se encontraban en las DTDs en XML 1.0 para especificar tipos de datos sobre elementos y atributos http://www.w3.org/TR/xmlschema-2/.

2. Metodología de descriptores para la recuperación de información en este ámbito

Como se ha comentado anteriormente, a la hora de recuperar información tanto para el proyecto OntoUAV o para la traducción de un documento especializado en el ámbito de los drones, es necesario definir una serie de descriptores lo suficientemente representativos para recuperar textos que contengan información relevante y útil, ya sea para traducir o para la construcción de la ontología. En este último caso, sugerimos trabajar la extracción de descriptores a partir de textos fiables, como son las normativas de países que hayan traspuesto a sus ordenamientos jurídicos el Reglamento (CE) No 216/2008 y Directiva 2009/48/CE para la creación de cursos de capacitación de pilotos de drones en España, ya que en el caso de la traducción se podría seguir la misma metodología a partir del texto original.

A continuación, nos disponemos a explicar los pasos de esta propuesta metodológica que podríamos dividir en:

a) Generación de términos índice: la normalización.
b) Creación de la sintaxis de búsqueda apropiada teniendo en cuenta los modelos de RI.
c) Búsqueda y evaluación de la documentación encontrada en Internet.

2.1. Generación de términos índice: la normalización

La generación de *términos índice* representa un avance en la documentación aplicada a la traducción pues en ocasiones el traductor

se encuentra con que no es capaz de generar una sintaxis de búsqueda apropiada para la documentación de un texto, proyecto o tema con el que trabaja asiduamente. La metodología aquí propuesta se basa en la metodología de Vilares (2005), aunque, en este caso la hemos aplicado al ámbito de la traducción, en concreto, al tema de los vehículos aéreos no tripulados.

Vilares (2005) indica que el proceso de generación de términos asociados a un documento se lleva a cabo a través de una serie de transformaciones en el texto. Este proceso, busca la simplificación de texto de "forma canónica" para posteriormente establecer una sintaxis de búsqueda apropiada, es lo que se conoce como normalización (*conflation*).

El proceso se divide en varios pasos, aunque algunos, como el *stemming* no nos servirán en este caso concreto (a menos que queramos hacer un tesauro, que no es lo que se pretende, ya que el *stemming* consiste en extraer la raíz del término.

Para realizar esta generación de términos índice será necesario disponer de un gestor de corpora[3] que nos permita realizar los pasos que a continuación vamos a describir, a saber, análisis del léxico, eliminación de las denominadas *stopwords*, eliminación de mayúsculas y signos ortográficos y la selección de los componentes de los documentos que se utilizarán por el SRI que llamaremos *términos índice*.

a) Análisis léxico del texto o *tokenización*

El principal objetivo de esta fase es la identificación de las palabras que componen el texto y determinar cuáles son más relevantes dependiendo de su frecuencia de uso y de su peso semántico. En primer lugar, convendría aclarar que el texto a analizar debe de estar en formato *.txt, para ello bastará con copiarlo y pegarlo en un archivo de texto plano y guardarlo en codificación UTF-8, para no encontrarnos con problemas de codificación de caracteres, sobre todo en francés.

Para realizar este análisis léxico utilizaremos el *tokenizador* de AntConc, en concreto la función *wordlist*, capaz de generar una lista

3. En este caso utilizaremos AntConc: un software de carácter gratuito que nos permite analizar textos y corpus de textos.

de todas las palabras que conforman el texto y clasificarlas por orden de frecuencia. Este programa nos dividirá el texto en palabras y las clasificará por orden de frecuencia.

Podemos considerar que estas palabras son candidatas a ser *términos índice* (Vilares 2005), no obstante, se han de efectuar una serie de modificaciones. Como hemos comentado previamente, AntConc genera el listado de todas las palabras del texto o del corpus ordenadas por frecuencia de uso, sin embargo, en esta lista además de los posibles *términos índice*, aparecen palabras con escaso contenido semántico (como pueden ser, artículos, preposiciones, conjunciones…), que suelen ser incluso más frecuentes en un texto y/o corpus que los *términos índice* candidatos, por este motivo, consideramos necesario eliminarlas tal y como explicaremos en el apartado b). El proceso de normalización tiene como principal objetivo la obtención de *términos índice* (simples o compuestos), con alto contenido semántico, que se repiten con mayor frecuencia y que son representativos del texto, es decir, los llamados KWIC o *key words in context.*

b) Eliminación de *stopwords*

Las *stopwords* (Kowalski, 1997) son aquellas palabras de escaso contenido semántico y por tanto de poca utilidad en este estudio debido a su excesiva frecuencia, como los verbos auxiliares o a su escaso contenido semántico, como artículos y preposiciones, por ello sería necesario eliminarlas de la lista de palabras. Para ello se creará en formato *.txt una lista de aquellas palabras que se desean excluir de la posible lista de términos índice, en este caso, convendría aclarar que en Internet se encuentran listados de *stopwords* en diferentes idiomas y, por ello, recomendamos la utilización de estos, ya que su elaboración de forma manual podría llevar bastante tiempo y, estos recursos nos permiten disponer de forma rápida y gratuita de ellos[4]. Por último, este archivo en formato *.txt se añadirá en el programa en las opciones del mismo para que las excluya al generar la lista de palabras, en el caso de AntConc se econtraría en *Tool Preferences>Wordlist>Word list range options>Use a stoplist bellow>add words from file.*

4. Como bien sabemos, tanto el traductor como el investigador suelen trabajar con la presión del tiempo, por lo que ahorrar en este sentido sería de gran utilidad.

Una vez introducido el archivo, si se vuelve a generar la lista de palabras, observaremos que ya no aparecen las que tienen escaso contenido semántico.

c) Eliminación de mayúsculas y signos ortográficos

Una vez realizados los dos pasos previos, es posible que la selección de palabras aún no sea del todo precisa, bien porque una misma palabra la diferencia por estar en mayúscula y en minúscula, considerándolas como términos diferentes, bien porque algunos signos ortográficos, sobre todo en el caso de los textos en francés, dificultan la lectura al programa del texto.

Para solucionar este problema habrá que modificar las opciones del programa, indicándole que trate todas las palabras como si estuvieran en minúscula. En el caso de AntConc habrá que recurrir a *Tool Preferences> Wordlist> Other options> Treat all data as lowercase.* Este paso es completamente necesario ya que de no hacerlo es posible que contabilice una palabra que aparece a veces en mayúscula a veces en minúscula como dos palabras en vez de una.

Por otro lado, será necesario que limpiar el texto de los signos ortográficos que den problemas, como pueden ser el acento circunflejo, las diéresis y tildes. Para la realización de este paso se recomienda la utilización de un editor de texto plano, como puede ser el Bloc de notas o el software Wordpad ++.

d) *Stemming*

El lenguaje humano tiene la capacidad de formular un mismo concepto de diferentes maneras, esto puede repercutir a la hora de realizar una comparación de documentos, ya que pueden denominar a un mismo concepto de diferentes maneras. Para reducir este problema "los sistemas de Recuperación Información recurren a la técnica de *stemming*" que consiste en la extracción de la raíz de las palabras, para luego realizar la búsqueda de la palabra y sus variantes.

Este primer paso para la recuperación de información no sería necesario, ya que se emplearía mucho más tiempo del que se ahorraría en una búsqueda manual. No olvidemos que en este proceso también nos preocupa la rapidez con la que se puede encontrar la información,

asimismo, nos encontramos ante un ámbito muy especializado en el que el uso de variantes es mucho más escaso que en una temática general.

e) Selección de términos índice

Finalmente, una vez realizado el proceso de *limpieza* y *tokenización*, nos encontramos con una lista de palabras ordenadas por frecuencia de uso en el texto.

La siguiente tarea consiste en seleccionar las 10 primeras palabras para realizar un estudio en profundidad de ellas, estudiaremos su empleo en *bigramas* y *trigramas* para observar si su empleo más frecuente es compuesto o simple. Para ello haremos uso principalmente de dos funciones disponibles en AntConc, la función *Concordance*, en la que se pueden observar las concordancias y las KWIC, es decir, cómo se usa ese término en el contexto del texto estudiado y la función *clusters*, que se encarga de ordenar las palabras con sus concordancias a la derecha y a la izquierda más frecuentes, para seleccionar el número de *grammas* que deseamos que aparezcan será necesario modificar las funciones de búsqueda de AntConc.

Una vez estudiados los términos, su frecuencia de uso como términos simples y como términos complejos, seleccionaremos aquellos términos más frecuentes y representativos del texto para continuar con el siguiente paso: la creación de una sintaxis de búsqueda a partir de los términos índice.

2.2. Creación de la sintaxis de búsqueda apropiada teniendo en cuenta los modelos de RI.

Para poder crear una sintaxis de búsqueda apropiada hemos de conocer, por un lado, el diseño de corpus que se desea compilar para la construcción de la OntoUAV y, por otro, cuáles son los modelos de RI, si no en detalle, al menos sí de manera general.

Bowker y Pearson (2002) realizan una propuesta de los criterios que deben incluirse a la hora de diseñar un corpus con fines específicos, como sería nuestro caso. Estos criterios serían: el tamaño, el número

de textos, el medio, el ámbito o tema, el tipo de texto, la autoría, el idioma/s, la fecha de publicación. A estos criterios, expuestos por Bowker y Pearson (2002) añadiríamos el tipo de corpus como criterio indispensable, además de los establecidos por Corpas (2001): tipo de textos contenidos en el corpus, especificidad de los documentos, cantidad de textos y codificación.

Por otro lado, en lo que respecta a los modelos de RI, los tres más utilizados y que poseen la mayoría de los buscadores generales que conocemos son:

a) Modelo booleano

El más conocido por ser el más simple, se basa en la teoría de conjuntos del Álgebra de Boole (Baeza-Yates, 1999). En este modelo el usuario formula una consulta firmada por una expresión booleana, con los operadores AND, OR y NOT. De este modo, el buscador devolverá al usuario aquellos documentos que considere relevantes, por ejemplo:

certificación AND drones

En este caso, el motor de búsqueda mostrará al usuario documentos que contengan estas dos palabras, la clasificación estará determinada por la relevancia que haya dado el buscador a los documentos dependiendo del sistema de clasificación que el buscador disponga, en el caso de Google, por ejemplo, sería el *pagerank*.

La principal dificultad que se le plantea al usuario sería la de saber formular su consulta, aunque una vez conocidos los términos claves y los operadores booleanos no debería ser complicado establecer una sintaxis de búsqueda.

b) Modelo vectorial

Pero cuando el usuario desea buscar documentos con palabras similares o grupos de palabras, en el caso que nos ocupa, el modelo booleano no es suficiente, por ello el modelo vectorial se plantea para dar solución a este planteamiento. Aquí, las consultas y los documentos son representados mediante vectores dentro de un espacio multidimen-

sional (Vilares, 2005) definido por los propios términos, y cada término definirá una dimensión.

Así pues, Vilares afirma que:

> El modelo vectorial no se limita, pues, a comprobar si los términos especificados en la consulta están o no presentes en el documento, como en el caso del modelo booleano, sino que la similitud entre ambos se calculan en base a los pesos de los términos involucrados, permitiendo de este modo, por un lado, la existencia de correspondencias parciales, y por otro, el cálculo de grados de similitud o relevancia conforme a los cuales los documentos pueden ser devueltos por orden de mayor a menor relevancia, facilitando notablemente el trabajo del usuario, que puede concentrar sus esfuerzos en los primeros documentos devueltos –aquellos más relevantes– o incluso definir umbrales de relevancia por debajo de los cuales un documento no es tenido en consideración.

c) Modelo probabilístico

Este modelo se basa en la Teoría del Modelo Probabilístico establecido por Robertson y Sparck Jones (1976). Se trata de calcular la probabilidad de que un documento sea relevante para la consulta que ha realizado el usuario, para ello el sistema suele tener en cuenta las consultas que se han realizado en unas condiciones similares (localización, idioma, frecuencia…) para proponer sus resultados e incluso proponer la sintaxis de búsqueda, como cuando el mismo buscador nos dice: "quizás usted quiso decir…".

Este modelo parte de dos afirmaciones:

1. Todo documento puede ser o no relevante para una determinada consulta.
2. El hecho de juzgar un documento dado como relevante o no relevante no aporta información alguna sobre la posible relevancia o no relevancia de otros documentos (Vilares, 2005).

De acuerdo con estos tres modelos y sabiendo cómo el motor de búsqueda podría responder a una consulta, el usuario tendrá los elementos necesarios para formular una consulta adecuada y que responda a sus necesidades[5].

5. No obstante, se recomienda consultar los operadores de búsqueda propios de cada motor de búsqueda para restringir la búsqueda lo más posible.

Por otro lado, para una mejor recuperación de la información o *best match*, sería muy recomendable que se creara una definición específica del tipo de documentos (DTD). En este caso, aconsejaríamos crear una DTD similar a la siguiente:

Drones.dtd

<!ELEMENT drones (dron+)>

<!ELEMENT dron (tipo, caracteristica, modelo, fabricante)>

<!ELEMENT tipo (#PCDATA)>

<!ELEMENT caracteristica (#PCDATA)>

<!ELEMENT modelo (#PCDATA)>

<!ELEMENT fabricante (#PCDATA)>

De este modo, los navegadores recuperarían mejor la información relativa a los drones y el usuario (empresario, traductor, terminólogo, fabricante, etc.), obtendría información fiable de manera rápida y eficaz.

2.3. Búsqueda y evaluación de la documentación encontrada en Internet

Para este último paso, el usuario deberá seleccionar el buscador que mejor se adapte a sus necesidades. Por regla general, el usuario suele recurrir a Google, aunque en ocasiones, cuando se encuentra ante una consulta muy especializada, sería necesario acudir a páginas específicas y no a un buscador general, como podrían ser las agencias estatales de aviación. En el caso de documentación sobre normativas (por ejemplo: traducción de textos jurídicos sobre este ámbito), o fabricantes de vehículos aéreos no tripulados, búsqueda de manuales para elaboración de glosarios técnicos o corpora técnico (en el caso de traducción de documentos técnicos). Nosotros nos vamos a centrar en la documentación sobre normativas, pues los textos de los que disponemos tienen esta tipología textual.

Si el usuario está familiarizado con Google, siempre puede utilizarlo como buscador principal, pero se aconseja, en la búsqueda especializada, restringir la búsqueda a estos sitios, para ello bastará añadir el operador *site:* y añadir la url en la que se desea buscar, por ejemplo, si el usuario desea buscar información en la web de EASA (European Agency por Security Aviation) podría realizar la búsqueda deseada en Google para que busque únicamente en EASA:

*drones site: w*ww.easa.europa.eu/

Por otro lado, conviene estar atento a los resultados propuestos por los buscadores, sobre todo cuando no conocemos las fuentes, para lo cual, aconsejamos utilizar la plantilla de evaluación de sitios web que propone la Universidad de Maryland[6].

3. Análisis y resultados

Una vez explicada la metodología a seguir, nos disponemos a mostrar el análisis y los resultados de la misma aplicada a los siguientes textos normativos tomados como modelo inicial para RI de OntoUAV.

- *The Swedish Transport Agency´s regulations on unmanned aircraft systems (UAS).*
- *Arrêté du 11 avril 2012 relatif à la conception des aéronefs civils qui circulent sans aucune personne à bord, aux conditions de leur emploi et sur les capacités requises des personnes qui les utilisent.*

3.1. Generación de términos índice: la normalización

Siguiendo la metodología explicada y, tomando como referencia los textos propuestos nos disponemos a seguir los pasos comentados

6. Se aconseja realizar este paso al principio, una vez el usuario se haya familiarizado con la evaluación de sitios web, no será necesario cumplimentar nada.

para la generación de términos índice en la realización del análisis, muestra de resultados y discusión.

a) Análisis léxico o *tokenización:*

En la tabla 1 presentamos el primer paso realizado en el estudio de la terminología de los textos arriba mencionados:

Tabla 1 – Análisis léxico preliminar

Orden	Freq.	Francés	Freq.	Inglés
1	1126	de	806	the
2	488	L	498	and
3	330	d	392	of
4	315	la	300	in
5	315	à	273	for
6	310	des	258	shall
7	294	le	254	be
8	291	et	225	to
9	283	les	207	a
10	258	du	204	The
11	223	un	176	flight
12	202	en	173	is
13	201	aéronef	168	Section
14	141	vol	156	aircraft
15	139	ou	150	or
16	133	au	149	maintenance
17	128	pour	139	that
18	123	télépiloté	120	with
19	117	une	109	s
20	106	aéronefs	108	which

Como bien se puede observar existe un alto número de palabras de escaso contenido semántico en ambas lenguas, tanto en francés “de, l’, d’, la, à…” como en inglés “the, and, of, for, in, for”. Cabe destacar que, sin haber aplicado ningún filtro, la primera palabra con contenido

semántico aparece en francés en el puesto número 13 "aéronef" y en inglés en el 11 "flight".

b) Eliminación de *stopwords,* normalización de mayúsculas y minúsculas y eliminación de caracteres innecesarios.

A continuación, presentamos la tabla 2 los resultados obtenidos tras la realización de la primera fase habiendo aplicado los filtros comentados en la metodología (*stopwords*, mayúsculas, minúsculas y caracteres innecesarios).:

Tabla 2 – Lista de palabras ordenadas por frecuencia de uso

Orden	Frec.	Francés	Frec.	Inglés
1	203	aéronef	186	Flight
2	143	vol	166	Aircraft
3	124	télépiloté	156	Maintenance
4	91	aéronefs	93	Company
5	82	civile	77	Uas
6	77	aviation	74	Pilot
7	77	télépilote	73	Manager
8	76	exploitant	72	Activities
9	70	chargé	71	Agency
10	70	ministre	71	Technical
11	66	télépilotés	70	Swedish
12	59	sécurité	69	Transport

En ella podemos observar una lista de frecuencias de las palabras de ambos textos de palabras con un alto contenido semántico.

c) Selección de términos índice

Una vez obtenida la lista de los términos más usados, nos hemos centrado en el estudio de los 7 primeros términos para comprobar si son términos compuestos o términos simples y observar cómo se utilizan en contexto.

Tabla 3 – Selección de los posibles términos índice

Orden	Frec.	Francés	Frec.	Inglés
1	203	aéronef	186	Flight
2	143	vol	166	Aircraft
3	124	télépiloté	156	Maintenance
4	91	aéronefs	93	Company
5	82	civile	77	Uas
6	77	aviation	74	Pilot
7	77	télépilote	73	Manager

Cada uno de los términos se ha estudiado en contexto, mediante la funcionalidad *concordance*, y los *clusters* a fin de observar las palabras que le acompañan y determinar si conviene considerarlo un término simple o compuesto.

A continuación, mostramos un ejemplo del término *aeronéf*.

Ilustración 7 – Wordlist o lista de palabras por frecuencia

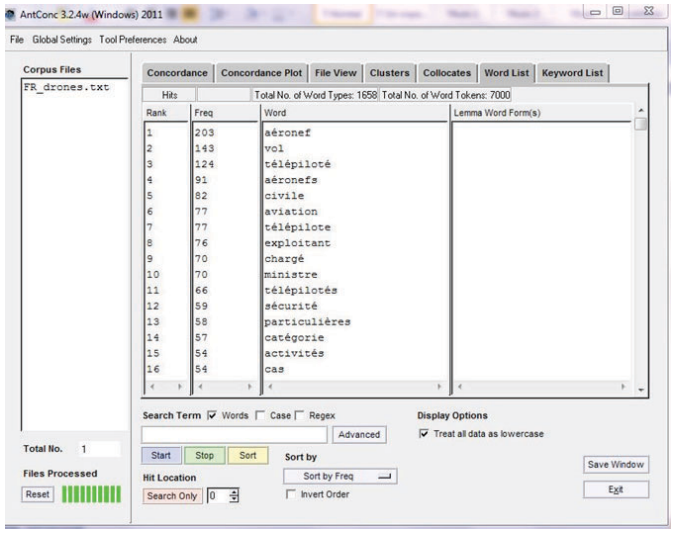
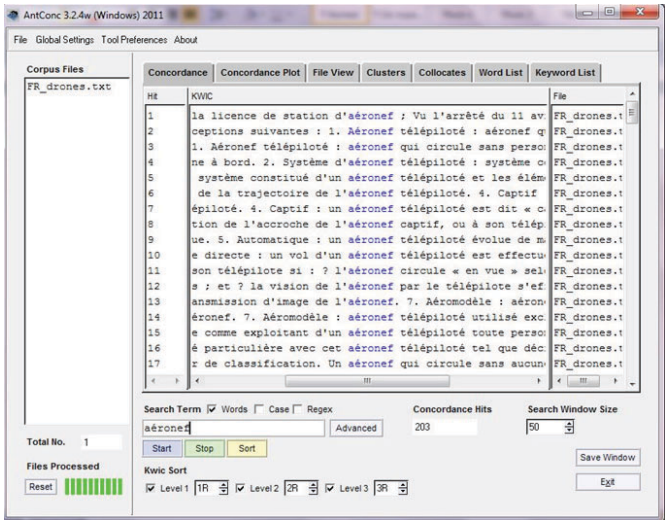
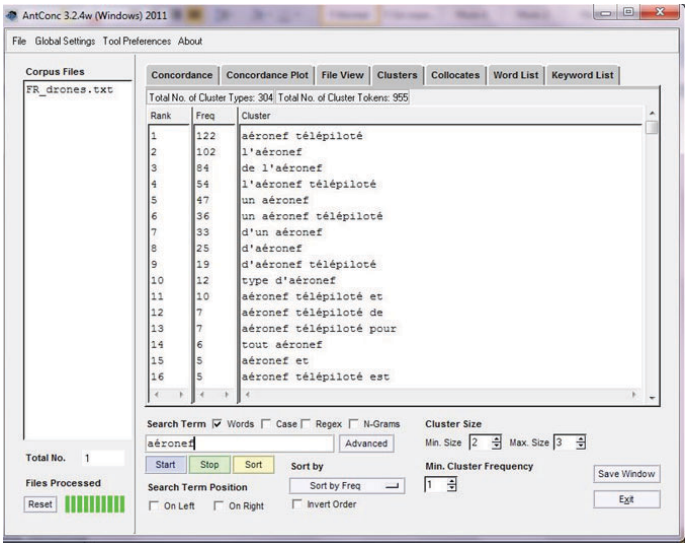


Ilustración 8 – Concordance o concordancias de ***aéronef*** en contexto



En la ilustración 8 ya se puede apreciar que *aéronef* aparece de forma repetitiva como *aéronef télepiloté.* Para asegurarnos, buscamos *aéronef* en *clusters* marcando como tamaño mínimo del *cluster* 2 gramas y máximo 3 gramas.

Ilustración 9 – ***Clusters*** con ***aéronef***



DELTA
35.4
2019

Como se aprecia en la ilustración 9 el bigrama *aéronef télepiloté* aparece 122 veces en el texto por lo que podríamos considerarlo como un término complejo y un posible descriptor.

Tras realizar este análisis con los términos seleccionados, pensamos que los términos índice que describen mejor los textos y, por tanto, el ámbito que aquí nos interesa son los que aparecen en la tabla 4.

Tabla 4 – Términos índice

Francés	Español	Inglés	Español
Aéronef télépiloté	Vehículo aéreo no tripulado Dron	Unmaned aircrafts system	Vehículo aéreo no tripulado Dron
Autorisation de vol	Autorización de vuelo	UAS	VANT
Aviation civil	Aviación civil	Fight	Vuelo
Sécurité aérienne	Seguridad aérea	Regulation	Ley Reglamento Regulación

3.2. Creación de la sintaxis de búsqueda: búsqueda y evaluación de la información

A partir de los términos índice seleccionados, correspondientes a las palabras más representativos de los textos, se construirán las diferentes sintaxis de búsqueda con la finalidad de compilar el corpus y/o documentación necesaria para realizar, en la OntoUAV o cualquier traducción relacionada con el ámbito de los drones. Para ello, además de la lista de palabras anteriormente presentada, se debe conocer la tipología textual que se desea buscar (por ejemplo, normativas o manuales técnicos), el tipo de encargo, en el caso de tratarse de una traducción, y finalidad ya sea de la traducción o de la RI. En este caso la OntoUAV, nos sirve para elegir el modelo de RI y el motor de búsqueda más apropiado. Es decir, el diseño del corpus debe ser claro, por ello se han de tener en cuenta los siguientes aspectos:

a) Para la construcción de la ontología hemos de compilar un corpus multilingüe, nosotros lo utilizamos en inglés, francés y español.

b) Partiremos de textos normativos y de manuales de pilotaje de drones, lo que nos asegurará la fiabilidad de los textos[7].
c) En los textos normativos, al conocer ya la existencia de normativas europeas, se utilizarán los descriptores para realizar por un lado, la búsqueda de un corpus paralelo multilingüe de normativa europea relativa a los vehículos aéreos no tripulados y, por otro, a la compilación de un corpus comparable multilingüe en las lenguas citadas.

Así pues, el diseño del corpus, tomando como referencia el modelo de diseño que proponen Bowker y Pearson (2002) y añadiendo como criterio el tipo de corpus (Corpas, 2001) quedaría de la siguiente manera:

Tabla 5 – Diseño de los corpora de la ontología

Tamaño del corpus	Aún desconocido
Nº de textos	Aún desconocido
Medio	Escrito
Ámbito/tema	Vehículos aéreos no tripulados
Tipología de textos	Normativos, manuales técnicos
Autoría	Administraciones estatales, empresas de fabricación de drones
Tipo de corpus	Multilingüe paralelo (textos UE) y comparable, textual, documentado
Lenguas	Inglés, francés y español
Fecha de publicación	A partir del año 2008, fecha de publicación del Reglamento (CE) No 216/2008

Una vez diseñado el corpus y los descriptores, podemos comenzar a construir la sintaxis de búsqueda. Por todas estas razones, hemos decidido hacer varios tipos de búsqueda combinando los diversos modelos de recuperación de información.

Además de los tipos de textos que se desean obtener, normativas y manuales, hemos de tener en cuenta que se buscará tanto un corpus paralelo como un corpus comparable, por lo que las sintaxis de búsqueda serán diferentes para cada caso.

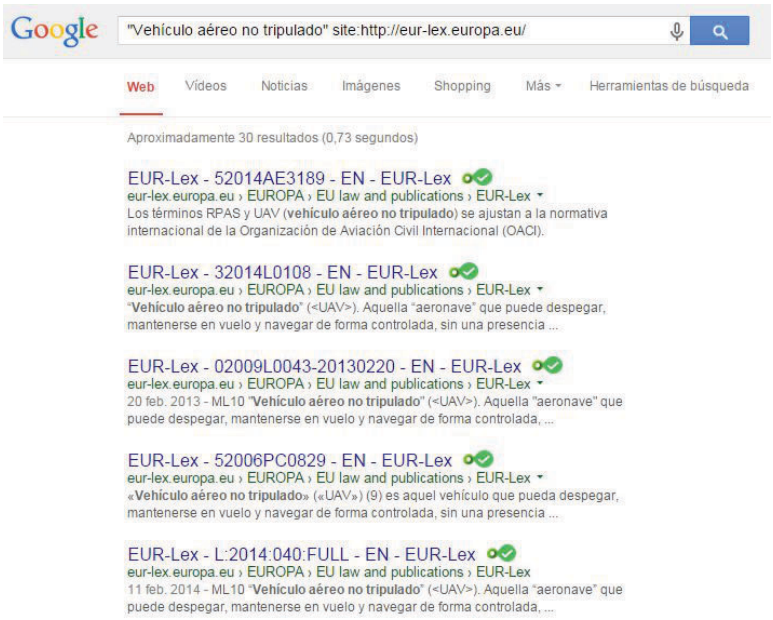
7. Para ello habrá que compilar dos corpora multilingües, uno sobre normativas y otro sobre manuales técnicos.

1. Normativas:

a) Búsqueda de normativa europea: se buscará directamente en el sitio de la UE destinado a la publicación de acuerdos y normas de la UE, Eurlex. Esta página permite la búsqueda multilingüe, lo que facilita la obtención del mismo texto en diferentes versiones lingüísticas, es decir, la búsqueda de corpus paralelo.
   a. si se utiliza directamente Eurlex, la sintaxis de búsqueda seleccionada sería la siguiente: "vehículo aéreo no tripulado".
   b. Si se utiliza Google: "vehículo aéreo no tripulado" site: site:http://eur-lex.europa.eu/. El operador *site:* le indica a Google en qué url buscar, de este modo no buscaría en todas las urls indexadas en el motor de búsqueda sino únicamente en Eurlex.

En ambos casos los resultados obtenidos hacen referencia a documentos legislativos de la UE. Si entramos en cualquiera de ellos, comprobaremos que tenemos todas las versiones lingüísticas, lo que facilita la búsqueda contextual de equivalentes.

Ilustración 10 – Búsqueda «vehículo aéreo no tipulado» site:http://eur-lex.europa.eu/



b) Búsqueda de un corpus comparable de normativas. Consideramos que la búsqueda se ha de realizar en repositorios institucionales de leyes y normativas o en boletines oficiales. En el caso del español se realizará la misma búsqueda en el BOE de las siguientes maneras:
   1) *«Vehículo aéreo no tripulado»* directamente en la página del BOE o "*vehículo aéreo no tripulado" site:* http://www.boe.es si se desea utilizar Google.
   2) Si se prefiere realizar una búsqueda más precisa utilizando el descriptor "ley", pues uno de los requisitos serían las normativas, se realizaría una sintaxis de búsqueda basada en el modelo booleano:

      «Vehículo aéreo no tripulado» ***AND*** *ley site:http://www.boe.es*

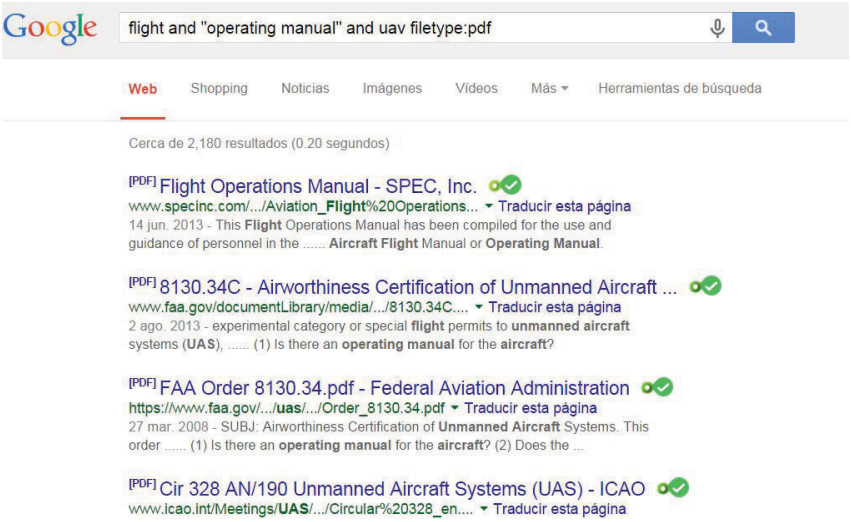
Esta búsqueda se podría realizar en la página *Légisfrance.fr* para Francia con el descriptor y/o descriptores "*aéronef télépiloté" **AND** loi,* en www.thegazette.co.uk para Reino Unido con el descriptor *"Unmaned aircrafts system" **AND** law **OR** act* y en www.irisoifigiuil.ie para Irlanda.

Siguiendo esta línea, se podrían realizar diferentes búsquedas utilizando los descriptores ya mencionados en las páginas de los boletines oficiales, siempre que se desee obtener información de las normativas vigentes.

2. Cuando utilizamos los manuales se ha de añadir un descriptor más a la lista y a la sintaxis de búsqueda: "manual" para español, "operating manual" e "instruction book" para inglés, "guide", "manuel d'utilisation", "instructions" para el francés. Además, se recomienda añadir a la sintaxis de búsqueda el operador de tipo de archivo *.pdf, para que quede delimitado, pues estos suelen presentarse de esta manera.

Mostramos un ejemplo para la siguiente sintaxis de búsqueda: flight and «operating manual» and uav filetype:pdf observándose los siguientes resultados.

Ilustración 11 – Sintaxis de búsqueda: flight and «operating manual» and uav filetype:pdf



Así pues, combinando los diferentes modelos de RI junto con los descriptores es posible obtener de manera empírica y rápida, documentación fiable y necesaria para compilar un corpus, ya sea para la creación de un recurso terminológico como es la ontología o para la traducción especializada.

4. Conclusiones

El establecimiento de una metodología para la RI aplicada a la traducción especializada y a la creación de recursos terminológicos y conceptuales es necesaria, ya no solo en el ámbito investigador que, en el caso de la construcción de la OntoUAV, facilitará la labor de búsqueda de documentos fiables, sino también en el ámbito profesional.

Esta metodología además de ayudar a filtrar aquella información que no es útil para los propósitos del corpus, supone un ahorro de tiempo, pues gracias a la extracción de descriptores aplicando un método automático y cuantitativo, no siendo necesario revisar el texto de manera manual para la búsqueda de estos. Solo se aplicará el método cualitativo una vez extraídos los posibles descriptores, lo que supone un ahorro considerable de tiempo.

Asimismo, realizando a una elección sistemática y objetiva de los descriptores, y teniendo en cuenta no solo el ámbito de aplicación sino también la tipología textual que se desea buscar es posible encontrar resultados fiables en pocas búsquedas, gracias a las técnicas de RI ofrecida por las ciencias de la documentación

Tras este estudio consideramos que es completamente viable el establecimiento de esta metodología para la extracción de descriptores en la creación de sintaxis de búsqueda y recuperación de información para la creación de una ontología multilingüe sobre vehículos aéreos no tripulados y, por tanto, para la traducción especializada en este ámbito. Dicha metodología ha respondido de manera directa e indirecta a los objetivos propuestos en este estudio:

1. Favorecer y facilitar el proceso traductor en el ámbito aeronáutico, ya que, como comentamos al principio, la traducción no puede concebirse sin documentación y, tratándose de un texto especializado sería prácticamente imposible traducirlo únicamente con el conocimiento lingüístico del traductor.

2. Favorecer la trasposición de la Directiva 2009/48/CE y el Reglamento (CE) No 216/2008 a los Estados miembros. Gracias a esta propuesta metodológica el traductor que en este ámbito aplique esta metodología será capaz de formular consultas que le proporcionen en un periodo breve de tiempo la documentación necesaria para sus traducciones.

Asimismo, completa los objetivos propuestos en el proyecto principal de OntoUAV, contribuyendo al procedimiento traductor con una metodología apta para cualquier tipo de texto especializado, siendo aplicable también a cualquier otro tipo de texto, conociendo sobre todo las fuentes de información propias del dominio de especialidad. También contribuye directamente al conocimiento del sector aeronáutico en materia lingüística, ya que a través de ella se puede recuperar información para la elaboración de repositorios documentales sobre el ámbito de estudio, diccionarios, tesauros y, sobre todo, para la elaboración de la OntoUAV.

Cabe destacar, como elemento a mejorar, que sería necesaria la creación de una DTD, como la que aquí proponemos o de un *XML*

Schema, para todas las páginas web relativas a los vehículos aéreos no tripulados, ya sean organismos oficiales, fabricantes o distribuidores, es decir, para aquellos sitios que dispongan de información fiable, cuyo objetivo no sería otro que el de indexar y situar correctamente y en una posición privilegiada sus páginas en navegadores web facilitando de este modo la recuperación de información.

Por último, añadir que la consecución de todos estos objetivos repercutirá de forma indirecta en el mercado aeronáutico ya que aumentarán las fuentes de información referentes al mismo y favorecerá la comunicación interlingüística entre agentes empresariales e institucionales.

5. Referencias

Abrirllave.com. 2014. *Documento XML asociado a una DTD interna | Tutorial de DTD | Abrirllave.com*. [online] Disponible en: http://www.abrirllave.com/dtd/documento-xml-asociado-a-una-dtd-interna.php [Consultado el 23 junio 2016].

ADAMIC, L.A., HUBERMAN, B. A. 2001. "Zipf's law and the Internet". *Glottometrics*, vol. 3, n 1, pp. 143-150. [En línea] Disponible en: http://www.arteuna.com/talleres/lab/ediciones/libreria/Glottometrics- zipf.pdf#page=148 [Consultado el 15 diciembre 2016].

_____. 2002. *Zipf's law and the Internet.*

ALTHUSSER, L. 2006. "Ideology and ideological state apparatuses (notes towards an investigation)". *The anthropology of the state: A reader*, 9(1), 86-98.

BAEZA-YATES, R. y RIBEIRO-NETO, B. 1999. "*Modern Information Retrieval*". ACM Press. Addison Wesley.

BOWKER, L. & PEARSON, J., 2002. *Working with Specialized Language*, Abingdon, UK: Taylor & Francis.

BRADFORD, S.C. 1934. "Sources of Information on Specific Subjects". *Engineering: An Illustrated Weekly* Journal. Londres, 137, enero, pp 85-86. Reimpreso como: Bradford, S.C. (1985): "Sources of information on specific subjects". J. Information Science, 10:4, 1985 (octubre)

BRODER, A. *et al*. 2000. "Graph structure in the web". *Computer networks*, vol. 33, no 1, pp. 309-320. [En línea] Disponible en: http://www.sciencedirect.com/science/article/pii/S1389128600000839 [Consultado el 15 diciembre 2016].

Codexexempla.org. *¿Qué es una DTD? < Curso acelerado < codexexempla.org*. [online] Disponible en: http://codexexempla.org/curso/curso_2_1.php#nota01 [Consultado el 23 junio 2016].

CORPAS PASTOR, G., 2001. "Compilación de un corpus ad hoc para la enseñanza de la traducción in versa especializada". *Trans*, 5, p.155-184.

CROFT, W.B. 1987. "Approaches to intelligent information retrieval." *Information Proccesing & Management*, 23, 4, p. 249-254.

DIRECTIVA 2009/48/CE del Parlamento Europeo y del Consejo de 18 de junio de 2009 sobre la seguridad de los juguetes. *Diario Oficial de la Unión Europea. L-170*, 30 de junio de 2009.

FOERSTER, H. 1949. *Cybernetics: Transactions of the Sixth Conference*, (editor), Josiah Macy Jr. Foundation: New York. 220 p.

FRAGA, A., MORATO, J. and SÁNCHEZ-CUADRADO, S. 2010. *DTD y SML Schema. Ingeniería de la Información*. [online] Universidad Carlos III de Madrid. Available at: http://ocw.uc3m.es/ingenieria-informatica/ingenieria-de-la-informacion/material-de-clase-1/03-DTD_y_XML_SCHEMA.pdf [Consultado el 23 junio 2016].

FRAKES, W.B. y BAEZA-YATES, R. 1992. *"Information Retrieval Data Structures and Algorithms"*. Prentice Hall.

Hipertexto.info. 2013. *DTDs y XML Esquema*. [online] Disponible en: http://www.hipertexto.info/documentos/dtds.htm [Consultado el 23 junio 2016].

JACQUEMIN, C. 1999. "Syntagmatic and paradigmatic representations of term variation". En *37th Annual Meeting of the Association for Computational Linguistics (ACL'99), Proceedings*, pages 341{348, Maryland.

_____. 2001. *Spotting and Discovering Terms through Natural Language Processsing*. The MIT Press, Cambridge, Massachusetts, USA.

_____ y TZOUKERMANN, E. 1999. "NLP for term variant extraction: synergy between morphology, lexicon and syntax". En Strzalkowski, T. (ed.). *Volume 7 of Text, Speech and Language Technology*. Kluwer Academic Publishers, Dordrecht/Boston/London, p. 25-74.

_____, KLAVANS, J. and TZOUKERMANN, E. 1997. "Expansion of multiword terms for indexing and retrieval using morphology and syntax". En *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the european Chapter of the Association for Computational Linguistics (ACL-EACL'97),* Barcelona, 7-10 July, p. 24-31, Madrid.

KORFHAGE, R. R. 1997. *Information Storage and Retrieval*. New York. Wiley Computer Publishing.

KOWALSKI, G. 1997. "Information Retrieval Systems: Theory and Implementation". *The Kluwer international series on Information Retrieval*. Kluwer Academic Publishers, Boston-Dordrecht-London.

MANDELBROT, B. 1968. 1965. "Information Theory and Psycholinguistics". En R.C. Oldfield and J.C. Marchall. *Language*. Penguin Books.

MARCOS ALDÓN, M. 2016. "Calidad y difusión de la documentación e información multilingüe en la web para traducción". En: S. Díaz Alarcón y E. Parra-Membrives (Eds.). *La traducción humanístico-literaria y otras traducciones* especializadas. Lit Verlag Dr. Hopf Berlin. p. 177-186.

MATURANA, H. y VARELA, F. 1995. *De maquinas y seres vivos*, Editorial Universitaria.

Mclibre.org. 2015. *DTD: Definición de Tipo de Documento. XML. Bartolomé Sintes Marco*. [En línea] Disponible en: http://www.mclibre.org/consultar/xml/lecciones/xml_dtd.html [Consultado el 23 junio 2016].

NAVARRO, G. y BAEZA-YATES, R. 1995. "*A language for queries on structure and contents of textual databases*". *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM Press, p. 93-101.

NEWMAN, M. E. J. 2005. "Power laws, Pareto distributions and Zipf's law". *Contemporary physics*, vol. 46, no 5, pp. 323-351. [En línea] Disponible en: http://arxiv.org/pdf/cond-mat/0412004.pdf?origin=publication_detail [Consultado el 15 diciembre 2016].

PARETO, V. 1945. *Manual de economía política*. Atalaya.

PEÑA, R.; BAEZA-YATES, R.; RODRIGUEZ, J.V. 2003. *Gestión Digital de la Información*. Alfaomega Grupo Editor.

REGLAMENTO (CE) No 216/2008 DEL PARLAMENTO EUROPEO Y DEL CONSEJO de 20 de febrero de 2008 sobre normas comunes en el ámbito de la aviación civil y por el que se crea una Agencia Europea de Seguridad Aérea, y se deroga la Directiva 91/670/CEE del Consejo, el Reglamento (CE) no 1592/2002 y la Directiva 2004/36/CE. *Diario Oficial de la Unión Europea. L-70,* 19 de marzo de 2008.

ROBERTSON, S.E. 1990. "On term selection for query expansion". *Journal of Documentation*, 46(4):359-364.

_____ and SPARCK JONES, K. 1976. "Relevance weighting of search terms." *Journal of the American Society for Information Sciences*, (27):129-146, May-June.

TOLOSA, G. y BORDIGNON, F. 2011. "Modelos y algoritmos de búsqueda + redes sociales para aplicaciones verticales de recuperación de información". WICC 2011. *Proceedings of XIII Workshop de Investigadores en Ciencias de la Computación*. pp. 243-248. ISBN: 978-950-673-892-1.

VILARES FERRO, J. 2005. *Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español*. Tesis Doctoral. Universidade da Coruña.

VON BERTALANFY, L. 1979. *Perspectivas en la Teoría General de Sistemas. Estudios científico-filosóficos*.

VON FOERSTER, H. 2002. *Understanding understanding*, volumen de papeles de von Foerster, publicó Springer-Verlag.

WANG, R. y KON, H. B. 1993. "Toward Total Data Quality Management (TDQM)". *Information Technology in Action: Trends and Perspectives.*

WIENER, N. 1949. *CIbernetics or control and communication in the animal and the machine*. The Massachussetts INstitute of Technology.

World Wide Web Technology Surveys. [En línea] Disponible en: http://w3techs.com/ [Consultado el 15 diciembre 2016].

YULE, G. 2007. *El lenguaje*. Ediciones Akal.

ZIPF, G.K. 1932. *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard University Press.

_____. 1935. *The psycho-biology of language*.

_____. 2013. *The psycho-biology of language: An introduction to dynamic philology*. Routledge.

Recebido em: 31/07/2017
Aprovado em: 10/09/2019